THE UNIVERSITY
of EDINBURGH

EVOLUTIONARY CONSERVATION AND DIVERSIFICATION OF COMPLEX SYNAPTIC FUNCTION IN HUMAN PROTEOME

MACIEJ PAJAK

Doctor of Philosophy
School of Informatics
University of Edinburgh

2017

Maciej Pajak:

Evolutionary conservation and diversification of complex synaptic function in human proteome

Doctor of Philosophy, 2017

SUPERVISORS:

Dr T. Ian Simpson

Prof Clive R. Bramham

ABSTRACT

The evolution of synapses from early proto-synaptic protein complexes in unicellular eukaryotes to sophisticated machines comprising thousands of proteins parallels the emergence of finely tuned synaptic plasticity, a molecular correlate for memory and learning.

Phenotypic change in organisms is ultimately the result of evolution of their genotype at the molecular level. Selection pressure is a measure of how changes in genome sequence that arise though naturally occurring processes in populations are fixed or eliminated in subsequent generations. Inferring phylogenetic information about proteins such as the variation of selection pressure across coding sequences can provide valuable information not only about the origin of proteins, but also the contribution of specific sites within proteins to their current roles within an organism. Recent evolutionary studies of synaptic proteins have generated attractive hypotheses about the emergence of finely-tuned regulatory mechanisms in the post-synaptic proteome related to learning, however, these analyses are relatively superficial.

In this thesis, I establish a scalable molecular phylogenetic modelling framework based on three new inference methodologies to investigate temporal and spatial aspects of selection pressure changes for the whole human proteome using protein orthologs from up to 68 taxa.

Temporal modelling of evolutionary selection pressure reveals informative features and patterns for the entire human proteome and identifies groups of proteins that share distinct diversification timelines. Multi-ontology enrichment analysis of these gene cohorts was used to aid biological interpretation, but these approaches are statistically under powered and do not capture a clear picture of the emergence of synaptic plasticity. Subsequent pathway-centric analysis of key synaptic pathways extends the interpretation of temporal data and allows for revision of previous hypotheses about the evolution of complex synaptic function. I proceed to integrate inferred selection pressure timeline information in the context of static protein-protein interaction data. A network analysis of the full human proteome reveals systematic patterns linking the temporal profile of proteins' evolution and their topological role in the interaction graph. These graphs were used to test a mechanistic hypothesis that proposed a

propagating diversification signal between interactors using the temporal modelling data and network analysis tools.

Finally, I analyse the data of amino-acid level spatial modelling of selection pressure events in *Arc*, one of the master regulators of synaptic plasticity, and its interactors for which detailed experimental data is available. I use the *Arc* interactome as an example to discuss episodic and localised diversifying selection pressure events in tightly coupled complexes of protein and showcase potential for a similar systematic analysis of larger complexes of proteins using a pathway-centric approach.

Through my work I revised our understanding of temporal evolutionary patterns that shaped contemporary synaptic function through profiling of emergence and refinement of proteins in multiple pathways of the nervous system. I also uncovered systematic effects linking dependencies between proteins with their active diversification, and hypothesised about their extension to domain level selection pressure events.

ACKNOWLEDGEMENTS

Writing this thesis would have not been possible without my supervisors - Ian, and Clive. Our weekly meetings gave plenty of opportunities to discuss research ideas, also, facilitated putting these ideas into action (and finally into written words). Ian was not only a helpful supervisor but also a great role model for research creativity, work ethic, and work-life balance. Clive stayed up to speed with my work even though we only communicated through Skype from time to time, which I find amazing. His insightful comments towards the end of the PhD program were crucial for determining the final shape of the thesis. Work presented in Chapter Five greatly benefited from a large amount of unique pre-publication insight kindly provided by Oleksii, a member of Clive's group. Also, Doug, who was part of my annual review panel, provided useful feedback which helped to steer my research in the right direction.

Besides people who provided academic advice, I would like to extend my thanks to people who helped me proofread the text - Andreas, Angus, and Kat. Also, Aga, and David provided great support in increasing throughput of jobs on the compute cluster. My research group in Edinburgh was a great crowd to go out with and spend long hours playing board games. Also, my second supervisor's group in Bergen welcomed me and made me feel at home and I will have very fond memories of my time spent hanging out with them. On a wider scale, the entire cohort of DTC Neuroinformatics provided stimulating and entertaining environment to work in. I hope EPSRC continues funding doctoral programs such as DTC/CDT (or any other permutation of these three letters).

On a more personal note, my girlfriend Emilia offered a tremendous amount of emotional support. Also, throughout my PhD I engaged in many activities which provided much needed distraction from work. From this place I would like to thank my dear friends at the Edinburgh University Wine Society, especially members of the Blind Tasting Team, also, fellow competitors at the Edinburgh University Ballroom Dancing Society, where my dance partner, Laura, deserves a special mention for putting up with me.

Last but not the least I would like to thank my family in Katowice and all my friends around the world for general support, patience, and understanding, especially in the difficult period of writing up when I often did not reply to their messages for prolonged periods of time.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, 2017

Maciej Pajak, March 6, 2018

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

aBSREL Adaptive Branch-Site Random Effects Likelihood

BF Bayes Factor

EBF Empirical Bayes Factor

FDR False Discovery Rate

FEL Fixed Effects Likelihood

GSEA Gene Set Enrichment Analysis

HMM Hidden Markov Model

IQR Inter-quartile Range

lncRNA long non-coding RNA

LRT Likelihood Ratio test

LTD Long Term Depression

LTP Long Term Potentiation

MAP Maximum a Posteriori

MCMC Markov Chain Monte Carlo

MEME Mixed Effects Model of Evolution

miRNA microRNA

ML Maximum Likelihood

MLE Maximum Likelihood Estimate

MRP Most Recent Positive Selection/Most recent branch with evidence for positive selection

MSA Multiple Sequence Alignment

ncRNA non-coding RNA

PSD post-synaptic density

PSP post-synaptic proteome

PTM post-translational modification

REL Random Effects Likelihood

1

INTRODUCTION

Natural scientists, philosophers, and theologists hypothesise about what makes humans special among other life forms on Earth (Gazzaniga, 2008). Out of all phenotypic traits of humans, researchers' attention is often drawn to our cognitive skills (Tomasello et al., 2005) Evolution explains that all our phenotypic traits, including cognitive abilities, must have been acquired in a process of a series of incremental changes started when the first forms of primitive organic life appeared on Earth.

1.1 MOTIVATION

The nervous system, just like any other system of a living organism, is ultimately a product of gene expression which starts in embryonic development and continues through the entire lifespan. Therefore genomic analysis is crucial to studying function and structure of the nervous system across all levels of analysis from molecular to behavioural.

We are only able to directly observe genetic sequence of contemporary taxa. Although partial information can be now extracted from ancient DNA (Hofreiter et al., 2001) it is generally limited by the natural decay of DNA strands outside of living cells so the scope of these studies is limited to relatively recent pre-historic human lineages, their contemporary animal species, as well as their gut biome. Attempts at extracting genetic information from older organic remains (e.g. insects enclosed in amber) suffers from the lack of replicability and is generally criticised (Pääbo et al., 2004).

As a result, we need computational inference to inform us about events which shaped our genome (and thus our phenotype) over time. When attempting to generate biological insights with computational tools we are facing a vast amount of data organised into different classes, each with their inherent complex structure - phylogenetic trees, mutation matrices, ontology graphs, protein-protein networks. All of them can be interpreted mathematically which allows for their manipulation and integration.

Research in this thesis is motivated by advancement in computational methodology and increasing availability of computational resources which now allow us to address open questions in the human genome's evolution and, more specifically, shed new light on the broad issue of the molecular origins of complex cognition.
The methodological prerequisite of such proteome-wide study is the ability to analyse multiple proteins in a consistent way, therefore modelling methods and analysis tools for modelling results form an integral part of the study.

1.2 BIOLOGICAL BACKGROUND

Having assembled methodological tools for modelling and results analysis; acquisition of modelling data for the entire human proteome brings an advantage of being able to infer systematic effects spanning multiple systems. It also provides background information when focussing on a single system and an opportunity for inter-system comparisons.
Patterns observed in the evolution of the full genome across large evolutionary distances have been the subject of recent research. For example, Harrison et al. (2002) studied emergence, growth, shrinkage and removal of protein families and found systematic effects such as power law distribution of family sizes and attributed large scale dying out of families to niche changes based on observations specific to bacteria. Family expansion was further studied by Lespinet et al. (2002), the authors focussed on early eukaryote taxa and found evidence for large-scale lineage specific family expansion. Further, Kurland et al. (2006) traced the origin of the eukaryote cell and its changes arguing that cell specialisation and compartmentalisation was a driving selective factor for eukaryote. Finally, Rands et al. (2014) attempts to link selection and function comparing coding and non-coding sequence within a modelling framework quantifying constrained genome using the measure of half-time of sequence based of pairwise comparisons between sequence from different taxa. The authors arrive at an estimate of 8.2% of the entire genome being under negative selection and likely to be functional.

In the remainder of this section I focus on a biological review of the molecular function in the nervous system.

1.2.1 *Neural signal propagation and processing*

Although the primary focus of this thesis is the molecular aspect of evolution, to fully understand its implications for the nervous system we need to understand the molecular setting within the basic building blocks of the nervous system. The principal cellular components of the nervous system are neurons and glial cells. Neurons form neural circuits which are responsible for electrical signalling and information processing; glial cells play a supportive role. Propagation of the electrical signal within a single neuron goes only as far as the length of its axon. Information processing potential is achieved thanks to the ability of neurons to connect, forming synapses - a spot where a terminal of the pre-synaptic neuron's axon meets the post-synaptic neuron, typically at a point located at one of its arborised dendrites branching out of this neuron's soma. Although a small proportion of synapses transfer electrical signal directly (electrical synapses), the majority of synapses forming the neural circuits in the nervous system are chemical synapses. In chemical synapses presynaptic terminal is separated from the postsynaptic one with a synaptic cleft and signal processing is achieved through a release of neurotransmitter from vesicles in the presynaptic side (Purves et al., 2012). A dense aggregation of protein complexes in the postsynaptic terminal closest to the synaptic cleft is often referred to as the post-synaptic density (PSD) (see Figure 1.2 for a microscopic image of a chemical synapse with clearly visible dense area of the PSD). Throughout this thesis PSD describes this part of the postsynaptic neuron, including proteins embedded in the postsynaptic membrane, whereas a closely related term - post-synaptic proteome (PSP) refers to the set of proteins which are present in PSD and detected in experimental studies such as Bayés et al. (2011). Detection of neurotransmitter by various receptors in the postsynaptic membrane triggers a variety of molecular reactions in the PSD. Their impact can be immediate, short term, or long term; they can affect the local environment, reach the nucleus of the cell or spread further. Out of all components of the nervous system the chemical synapses, and especially their post-synaptic part are of particular interest here thanks to their remarkable ability for fine regulation and lifelong plasticity (Kennedy, 2000).

Basic mechanism of synaptic transmission at a chemical synapse is outlined in Figure 1.1.

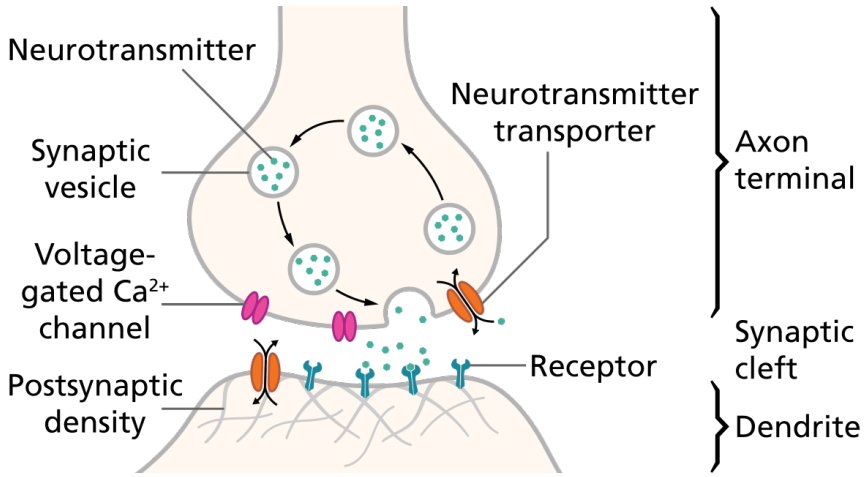


Figure 1.1: Synaptic transmission in a chemical synapse (diagram by T. Splettstoesser, 2015)

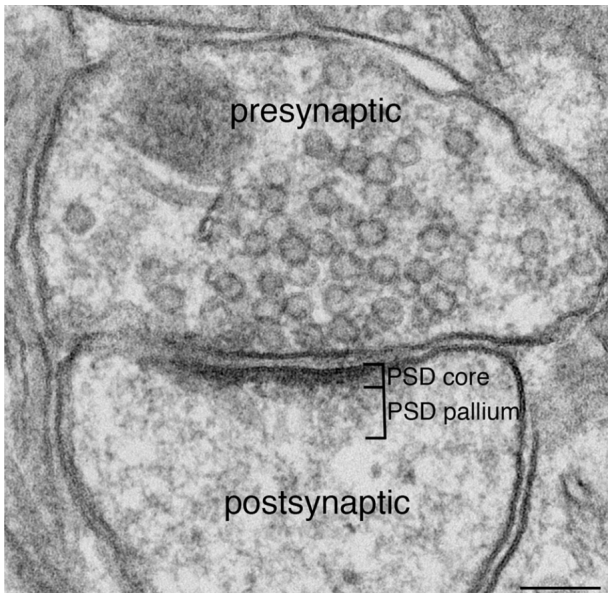


Figure 1.2: Chemical synapse, microscopic image, from Dosemeci et al. (2016)

The following sections outline how circuits comprised of neurons connected by synapses are first created, then used for signal processing, and modified in a flexible way through the life of an organism.

1.2.2 *Neural development*

Following the differentiation of stem cells to neurons the initial development of a network structure of a circuit is controlled by axon development and synaptogenesis. Axon guidance molecular signalling pathway directs the growing axon through

other tissue. A growth cone of the axon is guided with the mechanisms of chemoattraction and chemorepulsion by various classes of molecules - *ephrins, semaphorins, netrins*,and *cadherins*. Each of the signalling molecules is associated with their own assembly and delivery processes, receptors on the growth cone, and finally, intracellular signalling cascades (Baier and Bonhoeffer, 1992). Inside the expanding axon the growth cone directional changes are facilitated by actin cytoskeleton, then elongation of the axon is achieved through microtubule cytoskeleton. The disruption of finely regulated chemoattraction/repulsion pathways leads to observable cognitive deficits, for example mutation in a single gene encoding *ROBO3* protein (a receptor for chemorepulsive ligands) leads to a very specific deficit characterised by limited horizontal movement of eyes (as well as more general motor control issues) (Engle, 2010).

Once an axon arrives to its destination other molecular cascades take over and control creation of synapses through the process of synaptogenesis, however, selected signalling molecules involved in axon guidance continue their role here. Synaptogenesis starts from a local recognition mediated by *cadherins* and *protocadherins*, then further adhesion molecules (including *ephrins*) are recruited and the differentiation of pre- and postsynaptic densities begins. Eventually, cytoskeletal elements for vesicle mediated transport in the synapse are recruited by pre- and postsynaptic adhesion proteins such as *neuregulin* and *neuroligin*; the binding of these proteins is a key factor in localisation of all neurotransmitter receptors and their associated intra-cellular partners into postsynaptic density. Also, voltage-gated ionotropic channels are recruited by another adhesion protein - *neurexin* (Waites et al., 2005).

At this stage neurotrophic factors may regulate processes of neural growth or shrinkage as well as synapse stabilisation or elimination specific to the structure. *TrkA* and *TrkB* receptors relay the signal of *Neural growth factor (NGF)* or *Brain-derived neurotrophic factor (BDNF)* respectively to the intra-cellular signalling cascades. Separate cascades consisting of second messengers and kinases promote cell survival, neurite outgrowth and activity dependent plasticity (Farinas et al., 1994). Neurotrophic factors affect different structures to a varying degree; sensory sympathetic ganglia are particularly dependent on *NGF* regulation at development, over-expression (or under-expression) of *NGF* causes enlargement (or shrinkage) of sympathetic ganglia in mice (Fariñas et al., 1998).

1.2.3 *Synaptic transmission*

1.2.3.1 *Key presynaptic pathways*

Sequence of events occurring in the pre-synaptic neuron is fundamental for synaptic function and is tightly regulated within bounds of a few molecular pathways.

Neurotransmitters are released from vesicles in the cytoplasm in the process of exocytosis. The full cycle starts from an endosome containing neurotransmitter; a vesicle buds out of it and travels towards the membrane where it gets docked, primed, and, once pre-synaptic action potential reaches the synapse, fused with the membrane releasing a load of the neurotransmitter into the synaptic cleft.
Exocytosis is immediately followed by endocytosis where a vesicle buds from the outside membrane of the cell similar to how it buds from the endosome. It is then uncoated and transported back to the endosome, ready to be reloaded with neurotransmitter and reused (Takamori et al., 2006; Südhof, 2006).

Proteins involved in exocytosis and endocytosis play very specific roles, for example detatching of a vesicle from the reserve pool held together by actin cytoskeleton is achieved through phosphorylation of *Synapsin* by *CAMK2* which causes it to dissociate from vescicles. On the endocytosis end of the cycle adaptor proteins - *AP-2*, *AP180*, *Epsin*, and *Amphiphysin* and others assemble structures out of *Clathirin* which initiate budding of the membrane. *Dynamin* plays a part in the separation of a nearly formed vesicle; then proteins such as *Hsc70* and *Auxilin* remove the *Clathirin* skeleton (Südhof, 2004). It is easy to imagine that the spatial and temporal charateristics of such mechanism can be easily affected by slight changes in each of proteins involved. In an experimental study Shupliakov et al. (1997) disrupted interaction between *Dynamin* and *Amphiphysin* by introducing excess binding domains of these proteins which resulted in depression of neurotransmitter release due to impaired vesicle endocytosis.
Neurotransmitter release cycle pathway forms the presynaptic side of the wider neurotransmitter signalling pathway which can be divided into separate pathways for different synapses which use different neurotransmitters, the major excitatory synapse being the glutamergic synapse, and the major inhibitory one being the GABA-ergic synapse.

Another aspect of this pathway's function which should not be overlooked is maintaining reserves of the neurotransmitter. Local recycling of neurotransmitter occurs, and their reserves are also replenished by locally synthesised neurotransmitter molecules. Enzymes and peptide precursors are synthesised in Golgi appratus in the cell soma and then transported inside small vesicles along microtubule tracks all the way to the destination, at the same time enzymes required for local synthesis are transported much more slowly outside the vesicles (Goldstein and Yang, 2000).
Similar as in the endocytosis example, here fine regulation is crucial too. For example, Samaco et al. (2009) investigated a link between neurotransmitter levels and *Mecp2* protein (mutated in Rett syndrome and involved in local synthesis of neurotransmitters in dopaminergic and noradrenergic synapses). Animal study using $Mecp2^{null}$ mice allowed the authors to quantify the effect of protein depletion on neurochemical phenotype in aminergic neurons as well as observe mice behaviour. This study is particularly interesting as it demonstrates impact of very specific molecular regulation of one pre-synaptic pathway on the high level cognitive-behavioural phenotype.

1.2.3.2 *Key postsynaptic pathways*

Membrane of the PSD is rich in various receptors binding neurotransmitters from the synaptic cleft. Ionotropic receptors allow for the flow of ions through the membrane which changes polarisation of the PSD which, in turn, may trigger the activation or inhibition of the post-synaptic neuron. This process accounts for the fundamental function of the synapse which is the transmission of the signal from one cell to another through either activation or inhibition. Together with the neurotransmitter release cycle described in the previous section, the operation of ionotropic receptors contributes to the function of the neurotransmitter signalling pathway (Kennedy, 2000).

However, another class of receptors in the PSD membrane, G-protein coupled receptors, are more important in the context of this thesis. They do not open cross-membrane channels, instead, binding of a neurotransmitter to its outside part activates G-protein on the inside of the membrane which in turn can activate a long cascade of events. The signalling cascade starts with the G-proteins targetting the effector proteins - usually enzymes, such as *Adenylyl cyclase*. These enzymes produce intracellular second messengers, for example *Adenylyl cyclase* produces *cAMP* (Pin, 2000). Second messengers target further effector proteins, in the case of *cAMP* it is *cAMP*-dependent protein kinase A (PKA) which can phosphorylate specific proteins;

importantly, one of them is *CREB* - a key transcription factor in synaptic regulation (Delghandi et al., 2005).

In fact kinases can form a longer chain of signalling where one kinase phosphorylates another. *Mitogen-activated protein kinase (MAPK)* is a special example of a kinase cascade member, it also phosphorylates transcription factors such as *CREB*, thus regulating gene expression and linking local events in the synapse with the global machinery in the nucleus in the cell (Waltereit and Weller, 2003; Thomas and Huganir, 2004).

Overall, outside the PSD setting described here, G-protein coupled receptors cover a wide range of functional roles in living organisms, they are involved in detection of odours and pheromones, and are often targeted by drugs Neves et al. (2002), their role in cancer development has been a subject of research too Dorsam and Gutkind (2007).

1.2.3.3 *Molecular mechanism of synaptic plasticity*

Synaptic connectivity between neurons is not hard-wired, instead, it is susceptible to dynamic changes. The strength of a synaptic connection can be modified temporarily by processes such as synaptic facilitation, augmentation, potentiation, and depression. These phenomena can be explained mechanistically through relatively simple physical and chemical tuning of pre-synaptic density pathways. For example, facilitation can be explained through the build-up of Ca^{2+} ions following repeated stimulation. Similarly, depression occurs due to progressive depletion of the neurotransmitter pool (Zucker and Regehr, 2002).

Mechanisms responsible for long-lasting changes in synaptic strength - Long Term Potentiation (LTP) and Long Term Depression (LTD), are considerably more complex and depend on long cascades of reactions. On a circuit level, studies of synaptic connections in hippocampus circuits led to a fundamental observation that synchronised activity of two connected neurons leads to strengthening of a synapse between them; a phenomenon often described as Hebbian learning (Magee and Johnston, 1997; Caporale and Dan, 2008). However, LTP can also be induced in absence of coincidental activity, due to a pattern of sustained presynaptic activation in specific neurons (Urban and Barrionuevo, 1996). Molecular explanation of the LTP process can be divided into 2 phases - early and late.

Early LTP The early phase of LTP makes use of resources available at the synapse. The basic mechanism at glumatergic synapses (the most common excitatory synapses) relies on NMDA receptors. They require two coincidental events to allow flow of ions: depolarisation of the postsynaptic membrane (postsynaptic neuron activation), and neurotransmitter binding (presynaptic terminal activation). If these criteria are fulfilled NMDA receptors allow for the flow of Ca^{2+} ions into the PSD. Acting as secondary messengers they target *Calmodulin kinase 2 (CAMK2)* and *Protein kinase C (PKC)*. These kinases phosphorylate further effector proteins which contribute to exocytosis of additional AMPA receptors into the membrane resulting in increased permeability of PSD under neurotransmitter release from the presynaptic terminal (Kauer et al., 1988; Soderling, 2000).
Although this is the basic cascade of events, they can be tuned further; for instance, Esteban et al. (2003) found that *PKA* facilitates insertion of AMPA receptors by phosphorylating two of the receptor's subunits.

Early LTD follows the same logic, but AMPA receptors are removed from the synaptic surface through endocytosis process very similar to the vesicle recycling procedure described in section 1.2.3 and involving the same effector proteins such as *Dynamin*.

Late LTP Ultimately, lasting structural changes anywhere in the cell, and in this case, at the PSD, require building material in the form of new protein molecules.
As discussed in the previous section, kinases such *PKA* and members of *MAPK* family, specifically *ERK1* and *ERK2* target and phosphorylate *CREB* (Waltereit and Weller, 2003). Signalling pathways responsible for the early phase of LTP such as *CAMK2* and *PKC* converge on *ERK* subfamily kinases thus providing a link between the early and late LTP (Kelleher et al., 2004). Involvement of *PKA* - a downstream effector of G-coupled protein receptors supplements the mechanism of LTP-dependent transcription with additional layer of fine regulation through sensitivity to other neurotransmitters such as *Dopamine* or *Norepipherine* which activate these receptors (Delghandi et al., 2005).
After mRNA is transcribed from the DNA inside the nucleus proteins are translated elsewhere in the cell, local production proteins allows for faster structural changes in reaction to stimulus. Thus, a seemingly basic cellular function of protein translation plays a fundamental role in synaptic function (Job and Eberwine, 2001).
It is a two way process where LTP and LTD regulate translation initiation and in turn

translated proteins support the process of synaptic change through receptor number increase or dendritic spine growth (Klann et al., 2004; Pfeiffer and Huber, 2006).
The relationship between proteins involved in local translation and the characteristics of LTP was studied through knockout studies. Deletion of suppressors of translation initiation such as *GCN2* reduced the stimulation threshold of late LTP, and phenotypically, hippocampal memory task performance was reduced (Costa-Mattioli et al., 2005). In another study Banko (2006) used a different translation initiation suppressor - *4E-BP2* and found enhanced LTD in knockout mice measured in an electrophysiological procedure.
Importantly, the level of regulation studied here is very fine, as the nervous systems of the knockout animals in these studies were still practically functional. However, there are also regulatory proteins which are necessary for late LTP without affecting the basic mechanism of early LTP. *Arc* is a perfect example of a master regulator, as its knockout was found to prevent late LTP and thus long memory formation of any kind with intact short term memory (early LTP) (Plath et al., 2006).

1.2.4 *Higher cognitive correlates of synaptic plasticity*

Although initially studied in the context of simple behavioural changes, now it is widely accepted that molecular mechanisms of LTP and LTD are also related to high-level cognitive phenomena (Grant, 2003). Relationship with complex memory formation is particularly well studied (Martin et al., 2000; Collingride and Bliss, 1993). Evidence for the key role of the hippocampus in memory formation and retrieval converges from the long-term synaptic plasticity studies mentioned in the previous section and anatomical-behavioural level studies of hippocampal ablation. The missing link between the molecular mechanisms and behavioural phenomena is a topic of ongoing research which resulted in many important findings in the recent years, such as the discovery of how hippocampal place cells and grid cells represent spatial information about the environment (Moser et al., 2008). Associative learning underpins emotional responses too; studies focussed on the amygdala revealed its role in fear conditioning and attention. On a molecular level these cognitive functions are achieved through the mechanisms of long lasting LTP and LTD (Gallagher and Hollandt, 1994; Rumpel et al., 2005).

Many of the knockout studies mentioned in the previous sections, in the context of the quantitative molecular effects, also provide supporting evidence for complex regulation of cognitive abilities by numerous proteins. Samaco et al. (2009) revealed th effect of presynaptic regulator *Mecp2* on mice behaviour, and Anagnostaras et al. (2003) found a link between certain classes of G-protein coupled receptors and cognitive dysfunction in mice through genetic ablation of muscarinic M1 receptors. Then, specifically focusing on the late phase of LTP/LTD, a study of Costa-Mattioli et al. (2005) illustrated the link between activity dependent local translation at synapses and memory task performance; and finally, Plath et al. (2006) experiment with *Arc* genetic ablation points to the important distinction between short term cognitive effects and any kind of lasting learning which depends on early and late LTP/LTD respectively.

1.2.5 *Evolutionary emergence of complex synaptic function - molecular perspective*

The basic molecular mechanisms of long term synaptic plasticity, which allow us to reason about complex regulation in human, were widely studied in simple organisms, such as Aplysia, in the seminal works of E. Kandel's group (Castellucci and Kandel, 1976; Pinsker et al., 1973; Kandel, 2012).
Going further, based on the study of proteomes of even simpler organisms, it was established that some of the key synapse protein classes were present in prokaryotes and early eukaryotes, much earlier compared to the emergence of nervous system as such (Emes et al., 2008; Kosik, 2009). Then, the first protosynapses evolved in unicellular eukaryotic creatures such as choanoflagellates and *Porifera* (Sakarya et al., 2007). In these early ancestors of a postsynaptic terminal we can already observe a basic complex of scaffolding, receptor and signalling proteins working together. Interestingly, NMDA and AMPA glutamate ion channels only appeared in slightly later species of *Cnidarian* (Richards et al., 2008; Sakarya et al., 2007), these are the elements of the synapse which are essential for early LTP (see section 1.2.3.3). The presynaptic proteome evolved independently from the postsynaptic one, its origins can be traced to the simple mechanisms of endo- and exocytosis which are universal between pre- and post-synaptic neurons as well as other cells in living organisms, hence less attention was paid to molecular changes in these processes in the context of complex synaptic function and its cognitive correlates (Emes and Grant, 2012). Focussing on more

recent divergence points in the tree of life, in the past 100 million years very strong purifying selection was observed in genes coding for synaptic proteins in rodent and primate lineages. The conserved functions between those two lineages included cognitive, social and motor functions, specifically learning and memory (Bayés et al., 2012). Also, the expansion of G-protein coupled receptors family is hypothesised to be an important factor in nervous system evolution (Krishnan and Schioth, 2015). Collectively, the findings suggest that evolution of the synapse proteome (especially postsynaptic proteome) is a fundamental factor contributing to the emergence of complex nervous systems and complex behaviour.

1.2.6 *Anatomical perspective on the evolution of nervous system*

Stepping away from the molecular level of the nervous system, the evolution of the brain has been studied on an anatomical level too. Generally the most frequently discussed correlates of increased cognitive function between different species involve connectivity Thivierge and Marcus (2007); Schenker et al. (2005). More specific comparative studies theorise about particular indicators which correct for the general increase in brain size. Hofman (2014) points to the increase in the radial column number in the cortex which is a response to the progressive cognitive specialisation of brain regions observed in primates and human. The authors also suggest that the white matter ratio correlates with a perceived increase in cognitive abilities in the primate lineage (Hofman, 2001) and they claim that a theoretical model of connectivity in a graph explains that connections need to grow faster than the number of cells to maintain optimal computational capabilities (Hofman, 2008). Then, research by Hänggi et al. (2014) contributes to the discussion by differentiating intra-hemispheric connectivity (which can be largely explained by total brain size) from inter-hemispheric connectivity (only 9% variance explained by brain size); the latter is associated more with cognitive specialisation. Importantly, as shown in section 1.2.2, axon growth, synaptogenesis, and synaptic homeostasis are all controlled by molecular processes; thus, their evolution must also be reflected on a molecular level.

1.3 HYPOTHESIS AND PROJECT GOALS

In summary, complex synaptic function responsible for learning, as well as higher cognitive phenomena, results from the coordinated and finely tuned activity of mul-

tiple protein pathways ranging from neural development and organisation through neurotransmitter signalling to activity-dependent regulation. Recent studies identified many of the proteins enabling these processes as deeply conserved, and point to orthologs of human proteins in much simpler organisms such as fly or worm, or even simpler eukaryotes which lack nervous system altogether.
However, as discussed above, existing studies of synaptic density evolution were not detailed enough and missed time-specific and site-specific effects. I hypothesise that a more detailed study using previously unavailable methods and on a previously unprecedented scale can address these gaps.
The broad goal of this research program is to improve our understanding of the molecular mechanisms underlying synaptic plasticity through a quantitative description of protein diversification in the PSD as well as in the entire human proteome.

In order to approach this exploratory research question the first objective is to use the most recent phylogenetic methodology to assemble a scalable workflow for evolutionary analysis of large groups of proteins, placing emphasis on consistency of analysis, which will allow us to draw systematic inter-protein comparisons.
Following that, I aim to apply the phylogenetic framework to the entire human proteome, and subsequently, analyse modelling results for biological insights.
First, I aim to investigate global characteristics of the PSP evolutionary diversification timeline, as well as its most relevant systems and pathways, in the context of the full human proteome evolution.
Second, I focus on how the diversification profile of post-synaptic proteins relates to their function expressed through their interaction patterns in the protein-protein interactome as well as their role in functional pathways. Here, I extend the work of Emes et al. (2008), not only looking at how early through the evolutionary history specific classes of proteins appeared, but also what happened to them between then and now.
The final objective is to investigate evolution at sub-protein level to study what evolutionary patterns drive development of high resolution functional features of selected individual proteins, especially in the context of their interactions with each other.

1.4 ORGANISATION OF THE THESIS

In this chapter I briefly summarise the current state of our understanding of complex synaptic function and its molecular evolutionary origins. This helps to set the biological background to my research and define methodological aims as well as biologically grounded goals for how application of the methods will provide novel insights.

In the second chapter I focus on the development of a methodological framework which I then use to generate modelling results for the following three chapters. I review the principles of phylogenetic inference and explain how my modelling pipeline applies them.
In the third chapter I look at the global picture of the temporal aspect of diversification in the whole human proteome as well as selectively in synaptic proteins. I confront findings from analysing my modelling results with multi-ontology gene set enrichment analysis in a search for large scale patterns in molecular evolution.
The fourth chapter continues with time-specific diversification modelling results and integrates these findings with network analysis of the post-synaptic interactome based on static protein-protein interactions data as well as pathway annotations. I study the relationship between topological properties of individual proteins as well as groups of them in the interactome and their evolutionary characteristics.
In the fifth chapter I focus on a smaller group of proteins and use other protein sequence, structural, and functional data, such as post-translation modifications (PTMs); binding domains; and secondary structure. Data integration allows me to describe evolutionary effects between pairs of interacting proteins in high spatial resolution.
In the final chapter I draw conclusions spanning the entire body of research introduced in the previous three chapters and summarise the contributions of my thesis to our understanding of synaptic density as well as proteome as a whole. I discuss methodological limitations of inference presented in the thesis and propose how this work could be extended in the future.
Finally, appendices include long tables and additional figures which provide further information to supplement findings presented in Chapters 2-5.

2

MODELLING WORKFLOW METHODOLOGY

Following an introduction of the biological setting of this project and formulation of open research questions in the previous chapter, in this chapter I provide a description of the theoretical framework driving the computational methodology of my study. I review currently available methods for each stage of the phylogenetic inference pipeline, then I proceed to describe the modelling workflow assembly, where I justify specific choices and discuss issues related to executing such pipeline at scale, i.e. for the full human proteome.

2.1 INTRODUCTION

2.1.1 *Evolution as a macroscopic phenomenon and a molecular process*

We are often concerned with observable and quantifiable changes in an organism's phenotype, and the scope of analysed phenotypic information which ranges from high level animal behavioural features to bacteria's responsiveness to drug treatment. However, evolution of the phenotype of organisms occurs due to the underlying evolution of their genotype on a molecular level which allows us to use the molecular phylogenetic tools to reason about causes for the phenotypic changes over large periods of time.

Genotype evolution is a result of single nucleotide changes called point mutations as well as larger scale genome reorganisation events such a gene duplication, whole genome duplication, or recombination. In a high level overview, the dominant theory of evolutionary selection claims that point mutation can occur randomly in genetic code, both in coding and non-coding regions. Such mutation introduces polymorphism in the population (multiple variants of a gene start to coexist). Due to random genetic drift levels of different variants fluctuate even in absence of selection pressure. Selection pressure influences frequency of a mutation, and ultimately decides its fate - fixation or elimination. The direction of change depends on the direction of selection

pressure, positive will gradually increase frequency whereas negative will decrease it. Speed of change depends on the magnitude of selection pressure (Nei and Kumar, 2000; Li, 1997).

An alternative explanation, the *neutral theory of molecular evolution* (Kimura, 1984), accounts fixation (and elimination) of mutations purely to random genetic drift yet it is not widely accepted as sufficient explanation for all observed effects. As such in my work I follow the former assumption that there are regions of genomic sequence which evolve under selection pressure, not purely due fluctuations caused by genetic drift, even if it is justifiable to assume large stretches of DNA evolve in a neutral way. This allows me to infer causality between changes in genomic sequence and circumstances in which they occurred.

It is more common to analyse the evolution of coding regions of DNA, mutations in non-coding regions are harder to track and selection pressure is not easily defined in a theoretical way which would allow modelling it. Many of the methods described in this chapter relate exclusively to coding sequences, however, some principles apply to non-coding sequence too.

Mutations in a form of 1:1 nucleotide swaps are not the only changes of genotype studied by phylogenetic methods. Other basic low-level events are insertions and deletions, in case of coding sequence it would be multiplies of three nucleotides which do not disrupt the reading frame. Higher-level large scale reorganisation of genome through gene duplication, recombination, etc. is outside of the scope of this thesis.

2.1.2 *Phylogenetic inference pipeline*

In a typical molecular phylogetic workflow (Figure 2.1) the objective is to infer the course of evolution of a chosen protein across a given taxonomic range. In modelling terms, given the datapoints which represent contemporary species' sequences (the leaves of the phylogenetic tree), the task is to infer the underlying divergence tree topology as well as data for the remaining non-leaf nodes of the tree (which represent sequences of ancestral species); also, we want to infer the most likely transition paths between nodes. The inference process starts with a search for ortholog proteins in other organisms. Subsequently all orthologs are aligned to each other using one of Multiple Sequence Alignment (MSA) algorithms.

Then, the evolution model as well as the phylogenetic tree are fitted to the align-

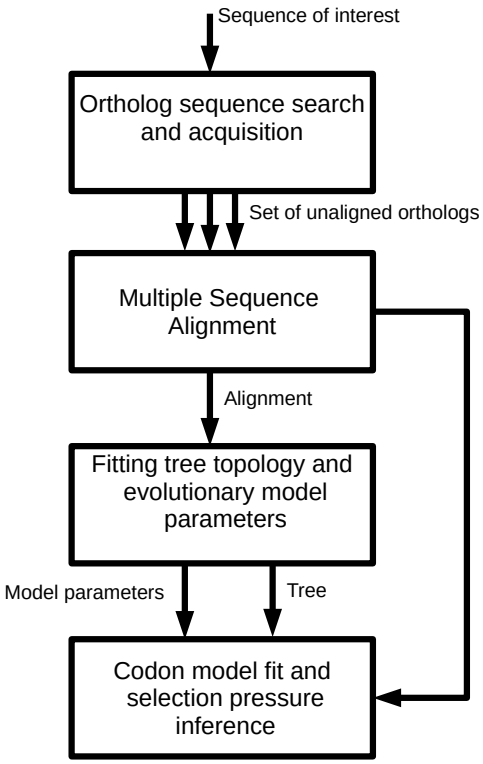


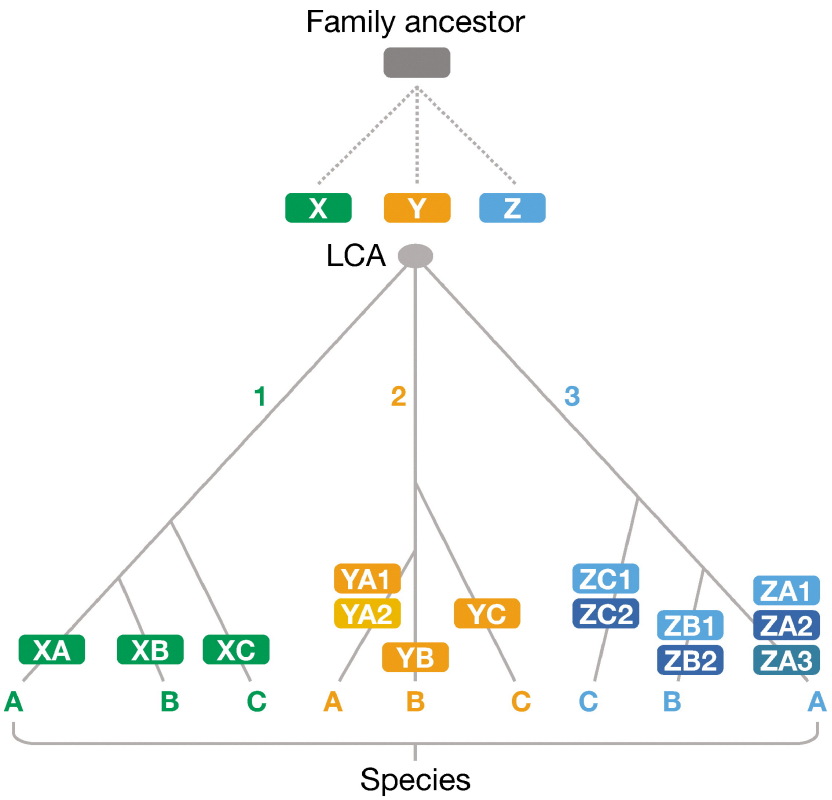
Figure 2.1: High level view of a typical phylogenetic inference pipeline

ment. Finally, we can use the alignment together with the model and the tree to infer selection pressure, either globally or locally. Importantly, at each of the stages we can manually intervene in the choices made by algorithms, which, however, requires some biological insight and can complicate streamlining the procedure.

A methodological overview of all stages is presented in the following sections.

2.1.3 *Ortholog search*

Having chosen a protein to analyse we can infer information about its evolution only if it can be compared to *similar* proteins in other organisms. The aim of the homolog search is to find these *similar* proteins. Two proteins are called *homologs* if they descend from a common ancestral protein and have not diverged far enough to lose sequence similarity. They often share the same name in closely related species but it might not be the case across more distant species hence we need to search for them

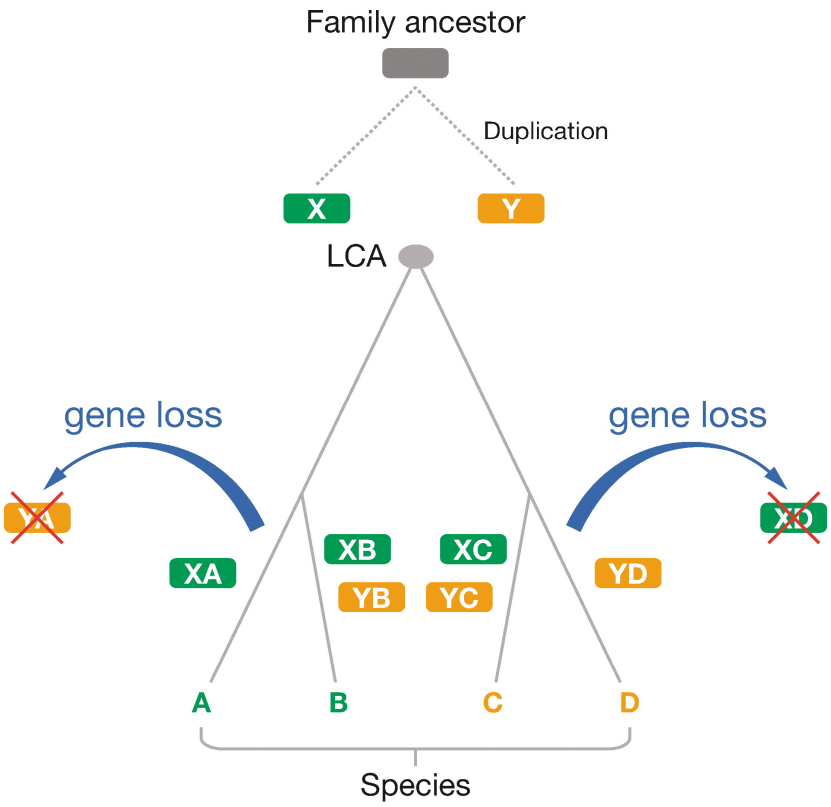


Koonin EV. 2005.
Annu. Rev. Genet. 39:309–38

Figure 2.2: Orthology and paralogy relationships for a gene family of 3 genes and 3 species. *XA*, *XB*, and *XC* are simple one-to-one orthologs (gene *X* in species *A*, *B*, *C*). Gene *Y* underwent another gene duplication event after speciation event defnining taxon *A*, these two paralogs *YA1* and *YA2* are both valid orthologs of *YB* and *YC*. Situation for gene *Z* is more complex, as each of the genes *ZB** is an ortholog of each of the genes *ZA** (similarly for *ZC**) - many-to-many orthology. From Koonin (2005).

using sequence information (normally amino acid sequence).

There are two distinct classes of homologs: paralogs and orthologs. Paralogs are homologs within the same taxon, they arise from divergent evolution of two duplicates of a gene. On the other hand, orthologs are homologs from different taxa, they arise as an effect of two species, which descend from a common ancestor, diverging over time Fitch (1970); Jensen (2001). Through the course of this project I am primarily concerned with orthologs as my aim is to highlight inter-species differences. Figures 2.2 and 2.3 illustrate more complex issues on the intersection of orthology and paralogy which come into play and may affect end results of the modelling framework even though they are often undetected (Koonin, 2005).



Koonin EV. 2005.
Annu. Rev. Genet. 39:309–38

Figure 2.3: Pseudoorthology due to gene loss. Genes XA and YD would normally be detected as orthologs while in fact they are paralogs, gene duplication occurred before speciation and subsequent gene loss. From Koonin (2005).

A common source of downstream error would be including false positives at this stage; on the other hand, missing orthologs can reduce the depth of the analysis, or introduce long gaps between divergence points where missing link taxa should have appeared.

BLAST (Basic Local Alignment Search Tool) is the most widely used procedure for homologue search (Altschul et al., 1997), however, numerous variants of the core algorithm exist (Altschul et al., 1997; Kent, 2002; Tatusova and Madden, 1999).

The search is conducted using short signature sections of the sequence and in case of a successful initial match, the length of compared sequences is extended. The procedure returns E value for each potential homologue which represents probability of getting the match by chance. In order to achieve good sensitivity a few runs with different parameters might be needed. Also, the list can be extended by additional searches run with already found homologs as input, since in principle homology relationship is transitive.

HMMer (Finn et al., 2011) is another homology search tool based on secondary structure of sequences and Hidden Markov Model (HMM) profile to inform search through the database. Although early implementations used to be much slower than BLAST, its execution time is now competitive with BLAST, and the approach is considered more accurate.

An alternative is using a database of pre-computed orthology results such as Ensembl Compara (Cunningham et al., 2015) where homology search results for each protein acquired in an well established pipeline are filtered and organised in a structured way (e.g. multiple ortholog hits in a taxon ranked by similarity, etc.). Ensembl Compara orthology pipeline starts from NCBI Blast+ (Camacho et al., 2009) search of every gene against every gene in all species, genes are then assembled into a sparse graph which is clustered with a hierarchical algorithm. In each cluster members are aligned (see the following section for common methodology of sequence alignment), and a phylogenetic tree is built with `TreeBeST`, pairwise relationships are confirmed after the cluster-specific tree is contrasted against the reference tree of life. The reference tree is a of phylogeny reconstructed as part of Taxonomy Project for all species catalogued in NCBI GenBank (Federhen, 2003).

Orthology relationships are further quality-checked through two pipelines - one focusing on gene order in the genome accounting for large scale reorganisations, the other based on whole genome alignment (see Vilella et al., 2009, for further details, and Ensembl Compara website for up-to-date refinements in the pipeline).

The main advantage is consistency across multiple proteins and potentially better functional validity of orthology relationship, however, putative orthologs in less well annotated taxa can be missed completely.

2.1.4 *Sequence alignment*

Having assembled a set of ortholog proteins from different taxa, in the following step we determine which sites correspond to each other across all the sequences in order to reveal the locations where evolutionary sequence changes such as mutations, insertions and deletions took place (Feng and Doolittle, 1987; Edgar and Batzoglou, 2006). This is the role of a sequence alignment algorithm. Here I describe how the problem can be solved optimally for two sequences, why the same principle fails with increasing number of sequences, and what solutions to this problem exist.

2.1.4.1 *Aligning two sequences*

When aligning two sequences we need to deal with substitutions as well as insertions and deletions, long repetitions can also complicate procedure by introducing ambiguity.

The most primitive, pre-computer era, methods of aligning two sequences are based on a dot-matrix where dots stand for a match between nucleotides/amino acids and diagonal lines of dots across the matrix represent an alignment.

The *Needleman-Wunsch* algorithm is a classic method in which two sequences are slid against each other and torn in some places in order to maximise the global score where matches between nucleotides/amino acids or likely substitutions are rewarded, whereas gaps and unlikely subsitutions are penalised (Needleman and Wunsch, 1970).

Typically the gap penalty is defined as

$$p = g + e(l - 1)$$

where p is the final penalty score, g is the penalty for opening the gap, e is the penalty for gap elongation and l is length of the gap.

Then, a substitution matrix specifies the score for aligning every possible pair of nucleotides or amino acids. These matrices are based on experimental studies and properties of nucleobases or amino acids. For example, when aligning nucleotide sequences Adeninine (A) and Guanine (G) are both purines, whereas Cytosine (C) and Thymine (T) are both pyrimidines, hence we get higher score for aligning A with G rather than A with C.

In case of amino acids preserving physicochemical properties such as polarity, hydrophobicity, or size is rewarded with a higher score.

When aligning two protein coding sequences it is a common procedure to align them at amino acid level and then map them back into nucleotide level sequences. Firstly, it preserves intact codons, second, matrices for amino acid substitutions are more precise than the ones for nucleotides because the nucleobase alphabet contains only 4 characters so, disregarding gaps, any two random sequences of nucleotides can be aligned with 25% identity. Most common matrices for amino acid sequences are PAM (Dayhoff and Schwartz, 1978) and BLOSUM (specifically BLOSUM62) (Henikoff and Henikoff, 1992).

Implementation of these constrains in the algorithm is based on Bellman Optimality

Principle, which states that a dynamic optimisation problem can be broken into a series of simpler subproblems and a subsolution of an optimal solution is optimal too. Practical application of this principle follows the logic that once the first subproblem's solution is optimised, then each subsequent subproblem is solved based on the state arising from the previous subproblem's (optimal) solution until the final subproblem is solved which grants a globally optimal solution (Bellman, 1957). In case of aligning sequences it means that we do not need to approach the whole alignment at once, instead we can build it gradually as long as in each step we find the optimal solution.

2.1.4.2 *Aligning multiple sequences*

A naive solution to the problem of aligning multiple sequences would be to align each one to all the remaining ones, however, this way computational complexity scales exponentially with the number of sequences which means that for long lists of sequences the procedure quickly becomes intractable.
Instead, various heuristics which allow to achieve polynomial complexity were developed. A few examples of widely used algorithms include `CLUSTAL-W` (Thompson et al., 1994), `T-Coffee` (Notredame et al., 2000), `ProbCons` (Do et al., 2005), and `Clustal Omega` (Sievers et al., 2011).
The common feature of these MSA implementations is progressive alignment. First, an approximate tree based on distance metric of sequence similarity is constructed with Neighbour Joining method (Saitou and Nei, 1987). Then, all sequences are progressively aligned to each other in a tree hierarchy, gradually aligning sequences and groups of sequences according to the hierarchical structure.
Furthermore, `T-Coffee` implements a consistency-based metric to improve its accuracy, `ProbCons` integrates HMM inference in the process, and `Clustal Omega` allows for iterative refinement of the alignment based on HMM profile - either an external one or inferred from the initial alignment. Importantly these methods tend to be robust to the preliminary tree choice (which *should not* be confused with the actual phylogenetic tree). All MSA algorithms can be benchmarked on a regularly updated BaliBase test set (Thompson et al., 2005). Apart from the benchmark score, the choice of the algorithm should be influenced by the computational cost and specific characteristics of the research question.

2.1.5 *Evolution model and its parameters*

As described earlier in this chapter, genotype evolution can be abstracted as a stochastic process where point mutations occur randomly and either become fixated or eliminated.
Substitutions observed when comparing aligned homologous sequences offer limited insight into the underlying process because through the course of evolution a nucleotide at a given position could have changed multiple times and having N sequences we can only observe N leaf nodes from the phylogenetic tree and we need infer the remaining $N-1$ nodes as well as transition paths between them.
Therefore we need to differentiate between true *genetic distance* and *observed distance*. *Genetic distance* between sequences on two nodes of a tree can be inferred with the likelihood function of distance as d which maximises $L(d)$ is a Maximum Likelihood Estimate (MLE) of the genetic distance (Nei, 1972, 1976).

Occurrence of mutations can be treated as a Poisson process (Cox and Isham, 1980). We assume four characteristics of this process: Markovian property, homogeinity, stationarity, and reversibility. For a time continous Markov chain transition probabilities can be derived from Kolmogorov relation:

$$\mathbf{P}(t) = exp(\mathbf{Q}t)$$

where Q is a relative transition rate matrix. For 4 nucleotides we can write the following 4×4 relative trasition rate matrix which represents relative rates of change of each pair of nucleotides:

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(...) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(...) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(...) \end{pmatrix}$$

It has 8 parameters $\pi_A, \pi_C, \pi_G, a, b, c, d, e$:

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

$$a + b + c + d + e + f = 1$$

$$rate_{site} \sim Gamma(\alpha, \alpha)$$

It is a relative rate matrix because we allow another parameter, rate, to vary across the sequence according to Gamma distribution with one free parameter.
The evolution model specifies the relationship between parameters in 4×4 $\mathbf{Q}$ matrix and constraints on them thus limiting the number of free parameters (Liò and Goldman, 1998).
For example, same as with sequence alignment, purine is more likely to mutate into purine and pyrimidine into pyrimidine hence we may want to differentiate between transistions (substitution within the family) and transversions (inter-family substitution) - thus we get at least 2 parameters.
The models vary in the number of parameters which need to be fitted. The simplest model, *JC69*, assumes equal rates of transitions and transversions and has no free parameters (Jukes and Cantor, 1969). As models become more complex constraints on parameters are gradually relaxed so that the most complex constrained model, *TN93*, has 6 free parameters in $\mathbf{Q}$ matrix (Tamura and Nei, 1993). Finally, if we eliminate all constraints between the 8 parameters we end up with *general time reversible* (GTR) model (Rodriguez et al., 1990) (also see the equation above).

At this point we are facing two optimisation problems, choice of the model and fit of model's parameters.
Normally choice of the model can be achieved with methods such as Akaike Information Criterion (AIC) (Akaike, 1974) or Bayes Information Criterion (BIC) (Schwarz, 1978). Both AIC and BIC formalise the trade off between complexity of the model and improvement on model's performance in explaining the data by asking the question whether the difference between the log likelihoods of the more complex model and the less complex model is big enough to justify choosing the complex model.
They are based on the likelihood function which for model selection has the form of $L(M, \theta, T) = P(D|M, \theta, T)$ where θ stands for a vector of model parameters and T stands for the phylogenetic tree (discussed in the following section). To simplify the likelihood function we would normally use MLEs of nuisance parameters (Sorenson, 1980), then the likelihood function becomes

$$L(M) = P(D|M, \hat{\theta}, \hat{T})$$

where $\hat{\theta}, \hat{T} = \arg\max_{\theta,T} L(\theta, T)$.
In an alternative (significantly more computationally demanding) approach, the likelihood values can be derived through Bayesian integration using a process such as Markov Chain Monte Carlo (MCMC) (Hastings, 1970). Then we obtain

$$L(M) = P(D|M) = \iint P(D|M, \theta, T)P(\theta, T|M)d\theta dT$$

When comparing models we can use Bayes Factor (BF), which is a ratio of model likelihoods obtained through integrating out any additional parameter vectors (Goodman, 1999). A simpler version would be Likelihood Ratio test (LRT), where MLEs of parameters are used instead. However, BF is preferred as it naturally guards against overfitting (using too complex model for the amount of data available), also, it achieves the correction for model complexity in an implicit way (unlike AIC and BIC).

Alternatively, with a longer list of orthologs we can opt for a more complex model without much risk of overfitting as there are more datapoints to base inference on. In that case we only fit MLE of parameters for a specific pre-determined model.
Importantly, in all cases we assume independence of sites hence likelihood of the full alignment is equal to the product of site likelihoods.

2.1.6 *Tree topology inference*

The terminal leafs of the tree are the homologue proteins which we previously aligned. In the next step of the analysis our task is to infer topology of a tree describing evolutionary relationships between the aligned sequences. In all methods we want to choose the tree that maximises an arbitrary function. This function can be based on parsimony, distance, likelihood or posterior probability. The important difference to fitting the evolutionary model is different treatment of parameters, we cannot nest tree topologies from the least complex to the most complex and we cannot order them by the number of free parameters. Also, we cannot integrate over a selection of tree topologies in the same way as we could do it with the evolution models (Nei and Kumar, 2000).

2.1.6.1 *Parsimony methods*

Using Ockham Razor principle (Ariew, 1976) we can search for the most parsimonous solution. In case of phylogenetic analysis the most parsimonous tree is the one that requires the fewest changes (substitutions/insertions/deletions) to generate the observed data. Naive implementation requires searching through entire tree space which is intractable because the number of trees grows factorially. For n homologs, there are $(2n-5)!!$ possible unrooted trees, and $(2n-3)!!$ rooted trees.
Conveniently, heuristics such as *nearest-neighbour interchange* (NNI), *subtree pruning and regrafting* (SPR) and *tree bisection and reconnection* (TBR) (Allen and Steel, 2001) allow us to search through fitness landscape more efficiently while reducing the risk of missing the global maximum. Their use is not limited to parsimony methods, on the contrary, they can be used in likelihood and Bayesian approaches too.

2.1.6.2 *Distance methods*

Given a matrix which summarises the difference between every two sequences with a single distance metric we can build a tree which connects similar genes with shorter paths.
The two simple algorithms which allow us to build a tree include unweighted pair group method with arithmetic mean (UPGMA) (Sokal and Sneath, 1963) and Neighbour Joining (Saitou and Nei, 1987) (mentioned earlier in the context of progressive alignment methods).
These methods are fast and easily scalable since they build a tree from scratch instead of searching through the tree space, however, they ignore a lot of information by reducing the information about sequences to a distance metric - a single number.

2.1.6.3 *Likelihood methods*

Maximum likelihood (ML) methods aim to chose the tree which maximises the likelihood function, i.e. $L(T) = P(D|T)$ where D is data - our sequence alignment and T is the tree topology. The value of the likelihood function is derived using the output of evolutionary model fitting - matrix $\mathbf{P}$. To be precise, the likelihood function is $L(T, \theta)$, where θ stands for the model and its parameters, however, implicitly we are using MLE of the model and its parameters (similar to the situation discussed in the section about evolution model fitting). In fact, it is a 'chicken and egg' problem whether the evolutionary model or the tree topology should be fitted to the data first, normally it

can be approached by iterative fitting of both.
Search through the tree space is normally initialised with a quick distance based estimate of the tree topology, then heuristics introduced earlier (NNI, SPR, TBR) are employed to expand the search space (Yang, 1997, 2007).

2.1.6.4 *Bayesian methods*

Bayesian approach goes a step further compared to the likelihood method and looks at the full posterior distribution of the tree given data, i.e. $P(T|D)$. This is where the MCMC method (Hastings, 1970; Metropolis et al., 1953; Chib and Greenberg, 1995) becomes particularly useful in searching the space of tree topologies to derive an estimate of their posterior probability. In order to speed up the process of sampling and thus convergence the standard practice is to improve mixing using Metropolic Coupling MCMC (MCMCMC) (Geyer, 1991, 1992).

2.1.6.5 *Rooting*

The objective of rooting is to introduce the last common ancestor of all analysed homologues to the phylogenetic tree. As it is an internal node, it needs to be inferred; when no information is available about its position in the tree, its location is determined computationally too. Two basic methods include *outgroup* rooting and *mid-point* rooting, both have their strengths and weaknesses (Smith, 1994). Mid-point rooting is quick and easy to implement, the root is defined as centre of mass of the branches - roughly equidistant from all leafs. In outgroup rooting we use prior biological insight about an ortholog in a distant taxon which definitely diverged from all others prior to the evolutionary events described by the rest of the tree. Although outgroup rooting is the most precise method it requires a careful and informed choice of the outgroup.

2.1.7 *Selection pressure*

As described earlier, mutations cause polymorphisms and their frequencies are affected by genetic drift and selection pressure (positive or negative). According to this conceptualisation the speed of fixating or eliminating mutation depends on the magnitude of selection pressure. However, genetic code is redundant and many mutations do not change the amino acid sequence of the resulting protein.
We define $\omega = \frac{\beta}{\alpha}$ as selection coefficient where α stands for synonymous (silent) sub-

stitution rate and β for non-synonymous (amino acid altering) substitution rate, in the literature alternative notation is often used: $\omega = \frac{dN}{dS}$.
$\beta < \alpha$ is associated with negative selection pressure and $\beta > \alpha$ positive selection pressure, the null/neutral selection pressure occurs for $\omega = 1$ (Yang et al., 2000). Selection pressure is inferred based on MSA, evolutionary model parameters and the tree topology derived in previous steps of the analysis pipeline.
First we fit a codon evolution model using Maximum Likelihood (ML)/Maximum a Posteriori (MAP) approach. The procedure is similar to the nucleotide model fitting, multiple alternative models exist which vary in their biological plausibility and complexity, measured by the number of free parameters to fit.

Common models include MG94 (Muse and Gaut, 1994) and GY94 (Goldman and Yang, 1994), in literature they are often reported in combination with a nucleotide level model which was used in earlier stages of the workflow, e.g. MG94 $\times$ HKY85.
Model can be chosen using the same criteria as for nucleotide models, i.e. AIC or BIC. Review studies generated some general guidelines about models' suitability, e.g. Shapiro et al. (2006) advocated using GY94 model (Goldman and Yang, 1994) based on AIC criterion. Another benchmark study concluded that some popular estimators of character distributions are biased, e.g. F3 $\times$ 4, therefore use of CF3 $\times$ 4 or MLF3 $\times$ 4 is advocated (Kosakovsky Pond et al., 2010).
Elements of a codon transition matrix Q for MG94 model are defined as following:

$$q_{ij} = \begin{cases} \alpha\theta_{mn}\pi_{np}, & \text{synonymous one nt } n \text{ to } m \\ \beta\theta_{mn}\pi_{np}, & \text{non-synonymous one nt } n \text{ to } m \\ 0, & nt\ distance(i,j) > 1 \\ -\sum_{k \neq i} q_{ik}, & i = j \end{cases}$$

Selection pressure can be computed at global level but it is more informative to look how it changes across the branches or across the codons of the sequence (Murrell et al., 2012).

The simplest and quickest solution is to use a counting algorithm. Numbers of synonymous and non-synonymous substitutions are counted averaging over all possible shortest paths of substitutions for each codon, then expected (neutral) proportion of

synonymous substitutions is calculated across all branches, subsequently we can test our codon-by-codon estimates of this proportion against the expected one using binomial distribution to determine if the difference is significant Kosakovsky Pond and Frost (2005).

More precise, yet more computationally demanding approaches use Random Effects Likelihood (REL) (Laird and Ware, 1982) or Fixed Effects Likelihood (FEL) (Kenward and Roger, 1997) models. Recently, a related approach based on multiple Hidden Markov Models was proposed by Mayrose et al. (2007).
Hybrid approach is also available where easily scalable counting algorithm is supported by more computationally demanding Bayesian model fitting (Lemey et al., 2012).
Finally, observation that selection pressure at any site varies through time inspired the most recent methodological advancements - Mixed Effects Model of Evolution (MEME) (Murrell et al., 2012), and Adaptive Branch-Site Random Effects Likelihood (aBSREL) (Smith et al., 2015).

2.1.7.1 *Site-specific selection pressure (FEL, REL)*

In FEL approach dN and dS rates are estimated directly at each codon - considered as a number of independent instances of the substitution process. Null model assumes $\mathrm{dN} = \mathrm{dS}$ and the alternative model allows dN and dS to vary, two nested models are compared with LRT to obtain a measure of confidence in the alternative model. FEL methodology only becomes accurate for a high number of sequences in the alignment (more than 20), generally REL is discussed as superior for smaller alignments (Kosakovsky Pond and Frost, 2005). In REL approach selection pressure is not estimated directly, instead it is considered a random effect distributed according to a discretised distribution. Parameters of the distribution are fitted to data according to MLE of codon evolution model parameters. Null model allows two bins of distribution - neutral, and negative, and alternative model allows the third bin - positive. These are nested models, hence for each site they can be compared through LRT.

2.1.7.2 *Branch-site selection pressure (MEME)*

MEME was introduced an improvement over FEL and REL which reconciles analysis of fixed-effect and random-effect within one framework (Murrell et al., 2012). The unique feature of this approach is the estimate Empirical Bayes Factor (EBF) in support

of positive selection model for individual codons on individual branches, however, authors discuss detection of specific isolated single branch-single site tuples under diversifying selection as not highly reliable due to small sample size per inference. Nevertheless, site-specific detection of positive selection is reliable and characterised by far superior power compared to FEL and REL regardless of the number of orthologs in the alignment according to authors' tests with simulated and real data. The method was introduced with GY94 $\times$ REV codon and nucleotide models and CF3 $\times$ 4 (which is corrected for the bias pointed out by Kosakovsky Pond et al. (2010)), however, other models can be used. Finally, LRT statistic for site-specific tests is distributed as a mixture of chi-squared distributions with degrees of freedom$\in \{0, 1, 2\}$.

2.1.7.3 *Branch-specific selection pressure (BSREL and aBSREL)*

aBSREL is an extension of the REL family of methods allowing model's complexity to vary between branches according to the information criterion. For each branch null model assumes a single rate of neutral/negative selection and the alternative model allows further rate distribution bins, adaptiveness of the model is expressed in an iterative selection of number of classes of rates.

For each branch, parameters for other branches are fixed, then rate class count is increased, parameters for that branch are optimised. The new model is contrasted against the previous one with AIC which results in either adoption of rate class number increase and another iteration of on the same branch, or rejection of the increase. In case of rejection procedure is repeated for the next branch, until complexity of alternative model for all branches is fixed.

The resulting model is not globally optimal, as at every step parameters for all remaining branches were fixed. At this point the model is optimised for all branches at once and the fully optimised model is the universal alternative model for positive selection testing (as well as for branch lengths). Finally, nested models are contrasted in a likelihood ratio test similar to FEL, REL, and MEME. Similar as for MEME, LRT statistic for site-specific tests is distributed as a mixture of chi-squared distributions with degrees of freedom$\in \{0, 1, 2\}$.

According to Smith et al. (2015) aBSREL achieves higher sensitivity over former branch random effects approaches due to overall decreased complexity of the alternative model, lowered computational complexity of fewer parameters being fitted is also an advantage.

2.1.8 *Manual intervention*

All steps of the pipeline can be altered manually, in some cases it may even be advisable to introduce manual corrections.

Ortholog search First of all, if the list of ortholog sequences is known a priori then the ortholog search step, and associated risk of false positives as well as false negatives, can be skipped. Also, we can manually filter a subset of sequences out of all search results according to the research question specific requirements (for a example a limited taxonomic range or minimum sequence similarity threshold).

MSA Although aligning sequences manually from scratch is intractable, if we want to focus on certain fragments of the sequence and we have prior knowledge about the alignment then we can manually edit alignment by sliding fragments of sequences, there is software available for it, e.g. Bioedit (Hall, 1999). Finally, if we are not satisfied with alignment of specific regions, or if we want to focus only on the selected fragments we can apply a mask to the alignment thus limiting the length of sequences.

Model selection and tree fitting Firstly, we can force choice of the model even if it is not the optimal one according to our selection criterion. In case we are using Bayesian fitting then we can also specify priors on parameters if we have some additional knowledge about them which is not encompassed by the model itself. Similar principle applies to codon evolution models.
Also, the tree topology (and even branch lengths) can be supplied a priori. In many cases the tree of life averaged over multiple proteins can provide an accurate topology after pruning down to taxa of interest. Also, a tree inferred from data can be filtered if spurious branch lengths or divergence points are noticed and there is biological rationale for assuming the inferred parameters are wrong.

2.2 WORKFLOW ASSEMBLY

Some of the most recent phylogenetic inference methods introduced in this chapter were widely applied in infectious diseases research first, e.g. a study of influenza virus by Rodrigue and Lartillot (2014) or a study of HIV by Pond et al. (2006). Large

scale human protein research tends to adopt new methodological approaches with a delay, therefore in my project I aim to bridge this gap and assemble a workflow with a selection of cutting edge methods which will be able to handle the entire human proteome. My modelling framework follows the high-level steps outlined in Figure 2.1. In the following sections I describe and justify decisions made at specific points of the workflow.

2.2.1 *Requirements*

First, considering the ultimate goal of applying the methodology with a view to gain novel insight into synaptic evolution, I compiled a list of requirements for my protocol.

1. Execution of the same modelling procedure for multiple inputs is an embarrassingly parallel problem, and I have access to Edinburgh University supercompute cluster - Eddie, thus ability to scale the execution of the protocol to this multi-node environment is crucial.

2. Further to that, an opportunity for speedup of modelling algorithms due to multiple CPU cores is welcome as this is available on Eddie.

3. From the practical point of view any methods which cannot be used through commandline interface are automatically excluded as automation is required at each stage of the inference.

4. Minimum human intervention along the way, it should be limited to organising batches of jobs to be submitted to the compute cluster.

5. On the methodological side, output of branch-specific selection pressure inference needs to be comparable between proteins.

6. In a similar way, site-specific selection pressure estimates for the full alignment need to mappable to specific amino acid locations in human sequence as well as other taxa sequences, this should also be ensured through consistency of sequence data between DNA, transcript and protein.

2.2.2 *Orthologs*

In this project instead of searching for orthologs with either BLAST or HMMer I opted for a database of pre-computed and quality-checked orthologs in Ensembl Compara (Vilella et al., 2009). The orthology relationship acquisition and filtering follows an established and continuously improved workflow which ensures confidence in results (see section 2.1.3 for more details). It is maintained by Ensembl Consortium (Cunningham et al., 2015) and updated regularly. Moreover, according to a comparative review Ensembl proves to be more reliable source of sequence information than RefSeq (Zhao and Zhang, 2015) which is important for satisfying requirements from the list in section 2.2.1.
The main advantage of this approach for my project was the ability to strictly control set of orthologs across all target proteins and confidence in biological relevance of homology relationships. Also, all taxa available in Ensembl Compara are mapped to a consensus tree of life available from the same source. Finally, acquisition of orthology information is easily automated programmatically. Results presented in this thesis come from Ensembl Compara version number 80 released in May 2015.

2.2.3 *Sequence acquisition*

Gene and transcript sequences were also sourced from Ensembl (Cunningham et al., 2015). For each taxon the single most similar (by sequence similarity, see Supplementary Table A.4) ortholog gene was picked. In some cases initially there were multiple matches per species listed, e.g. human *MAPK1* matches to mouse *MAPK1* and *MAPK3* but in my workflow mouse *MAPK1* is selected on the basis of similarity to the human sequence. In some cases this situation may represent many-to-many orthology relationships (see Figure 2.2), then it is not clear which paralog in a given taxon should be matched to preserve the functional relationship. Another drawback of this method is a possibility of including pseudoorthologs (see Figure 2.3). Despite the possibility of introducing noise in the form inaccurate ortholog choice a simple similarity criterion ensures general consistency of this step of the workflow.

Further, multiple transcripts for each ortholog gene were cross-referenced with Uniprot reference amino acid sequence of the protein coded by the gene to pick a

single protein-coding transcript (Bateman et al., 2015). The same transcript selection procedure followed for human transcripts.

2.2.4 *Sequence alignment*

Following available benchmarks (Pais et al., 2014) accuracy of the newest methods such as `MAFFT` (Katoh and Toh, 2010), `T-Coffee` (Notredame et al., 2000) and `Clustal Omega` (Sievers et al., 2011) is comparable. All these methods are all in active development, and are easily automated with commandline options; also, both `MAFFT` and `Clustal Omega` benefit from iterative refinement of alignment. `T-Coffee` requires substantially more computational resources (both time and memory) than the competitors while not offering better overall accuracy. Authors of the review point out good performance of `Clustal Omega` on poorly conserved termini of sequences which is a likely occurrence when using distant orthologs, hence `Clustal Omega` was selected as the MSA algorithm in my project (version 1.2.1 released in February 2014). Program was set for 2 iterations of HMM alignment correction, using first alignment to create HMM profile (i.e. no explicit HMM input). We used protein sequences translated from transcripts to create alignment as aligning amino acids is more accurate than nucleotides, then gaps were mapped back to nucleotide (codon) sequences.
Here I did not apply any masking of rows or columns of the multiple sequence alignment. A module of `T-Coffee` software allows for column-wise scoring of reliability of sequence alignment using *TCS* metric (Chang et al., 2014). Preliminary observations of the scoring module's behaviour revealed its sensitivity to highly variable regions of the sequence. While I appreciate that aligning them is often ambiguous and might introduce noise, they are also particularly interesting for the downstream analysis of selection pressure. Importantly elimination of such highly variable columns in a systematic way would have biased baseline, sequence-wide statistics which are used in assessment of the alternative model implying selection acting on specific sites.

2.2.5 *Phylogenetic tree and model fitting*

In order to be able to compare timelines of branch-specific selection pressure between different proteins we needed to use common phylogenetic tree topology. Fixed topology was based on the averaged tree of life (Cunningham et al., 2015), however,

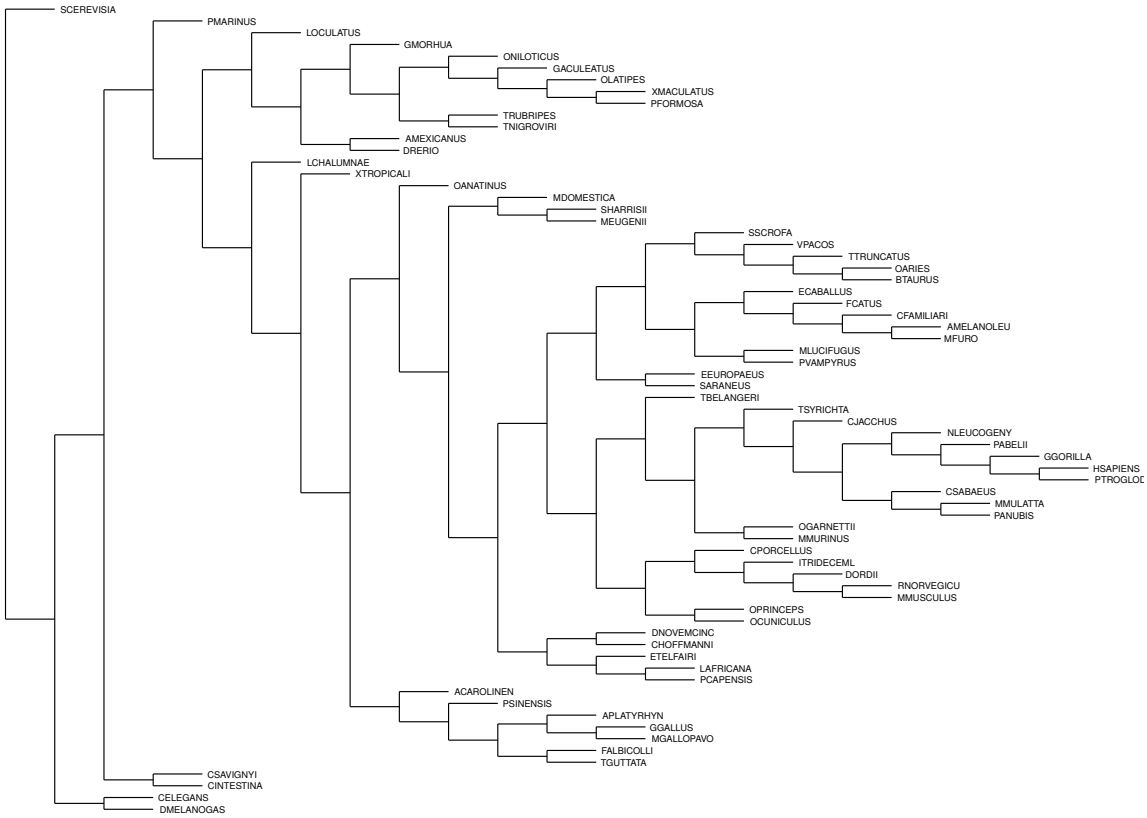


Figure 2.4: Tree of life topology from Ensembl Compara based on Taxonomy Project of NCBI GenBank (Federhen, 2003). Also see Supplementary Table A.3 for full names of all taxa used.

branch lengths were fitted individually for each protein in the evolutionary model fitting phase.

Large sample size (typically more than 30 transcripts in an alignment) allowed use of the most complex model of evolution (GTR/REV) without risk of overfitting (see section 2.1.5).

This step was completed with `HyPhy` software as part of selection pressure inference procedure described in the following section.

2.2.6 *Selection pressure inference*

My analysis is focussed on branch, site and branch-site models because in protein-coding sequences whole sequence analysis rarely indicates positive selection for two reasons: (1) majority of the sequence could be under purifying selection while only a

small region may be positively selected, (2) diversifying selection may occur in bursts which are brief in comparison to the full phylogenetic tree (Murrell et al., 2012, 2015). Therefore my modelling workflow uses branch, site and branch-site models which allow to explore fine-grained temporal and spatial patterns of diversifying selection separately, as well as a combination of both, which circumvents the risk of missing highly localised or transient signature.

Selection pressure analysis was completed with `HyPhy` package (version 2.2.4) (Kosakovsky Pond and Frost, 2005) controlled with extensive R and bash scripting to customise and automate the procedure and increase throughput. `HyPhy` is a software package which allows all kinds of selection pressure analysis outlined in section 2.1.7. It supports parallel computation through message passing interface (MPI) (Gabriel et al., 2004). Users can generate their own batch scripts for customised models and analyses. Authors also provide a remote server for executing the algorithm under default settings - `DataMonkey` (Pond and Frost, 2005) which however was not useful in case of my project.
The main competitor to `HyPhy` is `PAML` (Yang, 2007, 1997) which offers fewer customisation options and limits the scope for parameter adjustment. Also, it supports only REL approach while `HyPhy` support multiple methods including FEL and MEME. Neither of the packages is adequately documented, they both require substantial amount of scripting around them to run any non-standard analysis.

Within `HyPhy` I used FEL and MEME for site-specific inference, (MEME also provided site-branch inference), and aBSREL for branch-by-branch analysis. Conveniently MEME uses the same assumptions as FEL and is discussed as compatible with FEL (Murrell et al., 2012), on the other hand aBSREL is rooted in REL approach but is the only implementation available which is fit for the purpose (see section 2.1.7 for more details about principles of these methodological paradigms). For all methods I used GY94 $\times$ REV codon and nucleotide models and CF3 $\times$ 4 character distribution as per recommendation of Kosakovsky Pond et al. (2010) based on benchmarks.

2.2.7 *Implementation*

The majority of the workflow was implemented in `R/Bioconductor`, with small parts in `(Bio)Python` and bash. External software modules: `Clustal Omega` and `HyPhy`, were

compiled from source (`C` and `C++`) according to authors' instructions with `gcc` in `unix` 64-bit environment.
Further details are available in the supplementary section A.

2.2.8 *Execution and speed testing*

Following requirements for the modelling protocol outlined in section 2.2.1 I leveraged access to Eddie, Edinburgh University HPC cluster in order to execute my workflow in parallel for all proteins divided into batches.
In preparation for executing this modelling framework at scale I wanted to optimise distribution of computational load for total time of execution of all methods for the full proteome. As mentioned in section 2.2.1 I have access to University's HPC resources in the form of Eddie supercomputing cluster; it is a busy service with multiple research groups continuously using compute time on Eddie's nodes which means job scheduling and queuing system is in place. Here, not only did I want to test how much speedup is achieved by increasing core count but also I aimed to optimise flow of jobs through the scheduler queue.

In principle there are two levels of parallelism to be exploited here. There are no dependencies between jobs run for different proteins so all proteins (or batches of them) can be modelled at the same time. Also, selection model fitting is implemented making use of MPI which means speed-up can be achieved for each protein if more cores are allocated to a single job.
So, in theory, if 1000 cores were available for execution at any time and selection model fitting algorithm scaled perfectly it would not make a difference whether I run 1000 1-core jobs, 250 4-core jobs, or 10 100-core jobs, which means there would be no benefit of exploiting MPI implementation of model fitting. However, neither of these assumptions is true, hence the need for the test below.

Only the pipeline stages responsible for sequence alignment, and 3 selection pressure inference methods were executed in massively parallel fashion, however, sequence alignment computation time was approximately 2 orders of magnitude faster than any of selection pressure inference algorithms for the same protein so time spent aligning sequences becomes irrelevant compared to the selection pressure model fitting.

In this short test speedup was measured for MEME selection pressure algorithm as it was the most computationally demanding paradigm of selection pressure modelling.

2.2.8.1 *Speedup of a single job*

I selected 6 proteins which represented 2 levels of alignment depth (deep) and 3 levels of alignment length (short, medium, long), see Table 2.1. I used two alternative implementations of methods - one compiled for use with OpenMP shared memory environment, another compiled with OpenMPI libraries for scattered execution, I varied core count - 1 and 4 for OpenMP; and 1,4, and 16 for OpenMPI.

Results of the test are summarised in Table 2.2. Increasing core count in OpenMP implementation of the method did not improve execution time. Interestingly, even using a single core OpenMPI implementation is more efficient than single core execution of OpenMP method, furthermore, increasing core count in OpenMPI implementation improves execution time, achieving speedup in case of all proteins with the only exception being the long-deep case - see a note about reliability in the following section.

Table 2.1: Proteins used for speed testing, depth refers to the depth of alignment (number of orthologs).

	length		depth	
protein	aa	category	orthologs	category
USP	1102	long	60	deep
MYO1E	1107	long	28	shallow
SMIM15	74	short	54	deep
WFDC10B	73	short	20	shallow
C12orf29	325	medium	56	deep
OR1S2	325	medium	28	shallow

2.2.8.2 *Other practical considerations*

Execution time of a single job is not the only factor for consideration.
Fault exposure with scattered multicore approach needs to be taken into account - jobs occasionally need restarts after they fail and multicore jobs are more susceptible to it. For example in this limited test 16-core MPI job for the first protein (*USP*) had to be restarted twice, and for the fifth protein (*C12orf29*) - once, jobs with lower core

Table 2.2: Execution time in seconds and average speedup for each case, compared do single core OpenMP, and single core OpenMPI.

protein	1-core MP	4-core MP	1-core MPI	4-core MPI	16-core MPI
USP	101,102	100,935	71,702	26,000	71,999
MYO1E	47,547	49,368	28,973	8,191	1,859
SMIM15	5,862	4,889	2,647	893	230
WFDC10B	5,018	4,090	2,290	567	173
C12orf29	42,054	35,924	25,897	8,190	1,724
OR1S2	17,815	12,522	8,830	2,775	627
speedup vs 1 MP	1	1.16	1.85	6.11	22.38
speedup vs 1 MPI	-	-	1	3.38	13.89

counts did not suffer from it.

Although OpenMPI supports various techniques improving fault tolerance, they are not exhaustive and perhaps do not cover all situations which happen on a busy HPC cluster. For example, since cores are scattered between separate nodes other jobs residing on these nodes and increasing overall load may cause large disproportions in time taken to return partial results from different cores. Finally, fault tolerance when running in a distributed mode is affected by the implementation of the algorithm as such, and it was beyond the scope of this thesis to attempt further optimisation and elimination of bugs in the source code of `HyPhy`.

Another non-negligible factor is allocation of execution cores to jobs in the queue dependent on scheduler characteristics, if scaling is close to linear for 4 and 16 cores then total number of cores that scheduler allows me to occupy at any time becomes more important. When executing a large batch of exclusively either single core, 4-core, or 16-core jobs at a similar time of the same day during the week (comparable overall load of the supercomputer), total number of cores in use at any point was the highest for the 4-core jobs (up to 1100 cores), followed by single core (650 cores), and lowest in the 16-core case (600 cores).

Overall, using MPI version of the algorithm with 4 cores was the optimal solution accounting for the flow of jobs through the queue and MPI fault susceptibility.

2.3 MODELLING RESULTS

The full modelling framework was applied to 18544 human proteins, as listed in Human Proteome Atlas (accessed Dec 2015) (Uhlen et al., 2015).
Overall, spatial results of FEL and MEME models were available for 18269 proteins (98.5% of all). Temporal results of aBSREL model were available for 17636 proteins (95.1% of all).
In order to eliminate dropouts due to unusually long alignments (extreme execution time which exceeded allocated time on compute cluster) or random node failures, initial dropout proteins were rerun with substantially higher execution time and safer environment (i.e. no scattered multi-core execution). Upon inspection of a selection of failed proteins I concluded that the reasons for remaining dropout were very low number of orthologs annotated or highly gapped alignments which were both outside of my control; both cases may cause the model to fail at fitting maximum likelihood parameters due to lack of convergence. Low number of orthologs affected aBSREL to a higher degree hence larger relative dropout rate.

2.4 DISCUSSION

Work described in this chapter was motivated by building a customised, scalable workflow for phylogenetic inference of thousands of proteins. The resulting modelling framework which I built ensures consistency between different proteins in aspects such as set of available taxa, source of sequence data, number of parameters of the evolution model, and tree topology.
Admittedly, prioritising consistency of analysis between proteins may compromise accuracy of analysis for a small number of proteins, however, this thesis is concerned primarily with large effects spanning multiple proteins and differences between them which otherwise could not be quantified reliably (e.g. had I used different models or different trees between proteins). The following three chapters describe different angles of analysis of modelling results.

2.4.1 *Other aspects of molecular evolution*

Apart from local deletions, insertions and substitutions gene evolution is affected by other factors such as recombination, lateral gene transfer, gene duplication, whole

genome duplication. They can affect accuracy of phylogenetic inference applied in a standardised fashion to the full proteome.

Gene duplication If a fragment of DNA containing an entire gene becomes duplicated the two genes start evolving completely independently which might result in them acquiring different domains and different functions. Duplicated genes can go through *pseudogenisation*, *subfunctionalisation*, *neofunctionalisation*, or they can conserve their original function (Zhang, 2003). Gene duplication has major role in evolution of species. It is possible to detect duplication events when we cross-reference phylogenetic tree constructed for a family of genes with the species tree (*tree of life*). An example of application of this method can be found in studies of the protein family of *globins* (Efstratiadis et al., 1980; Shen et al., 1981). Generally, it should not affect accuracy of phylogenetic inference in this thesis, in certain cases multiple proteins in the same family may share elements of the tree close to the root (before the family expanded from a common ancestor).

Lateral gene transfer (LGT) Also known as horizontal gene transfer (HGT), as opposed to regular vertical gene transfer (from parents to offspring), this term describes transfer of genes between species.
It occurs frequently in bacteria and allows them to evolve antibiotic resistance (Gyles and Boerlin, 2014). However, traces of this process can be found in eukaryotes too, for example, ferns acquired their chimeric photoreceptors from another plant - hornworts (Li et al., 2014). It can be detected by comparing phylogenetic tree of one gene in an organism to other genes in the same organism, if it is drastically different from a certain point but at the same time resembles pattern displayed by a different organism then we can treat it as a candidate for LGT. Also, graph operation of *subtree pruning and regrafting* (mentioned earlier in the context of exploring the tree space) can be used to model the process of LGT (Allen and Steel, 2001).

Recombination It occurs when two different strands of DNA combine into one. The process occurs naturally during meiosis in diploid organisms, its function is to shuffle alleles in the next generation. If such event occurred within one gene, tracing its evolution with a single phylogenetic tree would be incorrect because in fact at a certain point there would be two separate trees describing evolution of two source DNA fragments (Schierup and Hein, 2000). There are multiple ways of computational detection

of recombination events, e.g. genetic algorithm by Kosakovsky Pond et al. (2006) or change-point process model by Minin et al. (2007). Overall, among all phenomena discussed here, undetected recombination events pose the highest risk of introducing noise into modelling results in this thesis (Posada and Crandall, 2002).

3

GLOBAL OBSERVATIONS FOR LARGE SETS OF PROTEINS

Chapter 2 reviews phylogenetic methodology background and outlines assembly of the modelling framework motivated by the broad project goals in mind (see section 1.3). My modelling pipeline uses an established framework of phylogenetic inference with the most recent algorithms integrated in it. It allows for selection pressure inference which informs about spatially or temporally limited diversification events for each protein. The modelling workflow was built, tested, and executed for the entire human proteome generating a complete dataset of probabilistically inferred evolutionary history of each protein through three complementary modelling paradigms - FEL, aBSREL, and MEME (see section 2.1.7).

In this chapter I use the modelling results generated by the pipeline from the previous chapter. I present how temporal information about episodic diversification periods can be aggregated across all proteins, this leads to protein clustering and extraction of informative features. Also, I discuss how emerging patterns and groupings can be interpreted by integrating data about divergence points as well as functional annotation from multiple ontologies.

3.1 INTRODUCTION

Exploratory analysis of large datasets often leads to identification of patterns which differentiate datapoints and allow us to reason about the underlying explanation for the groupings. In the domain of bioinformatics a common approach is investigation of the relationship between commonalities in experimental/modelling features and established functional data about what the datapoints represent, e.g. regions of the genome, proteins, chemical compounds, tissue sources, etc.

3.1.1 *Clustering and feature transformations*

A natural first step when aiming to explore structure in a dataset is to cluster datapoints. The outcome of clustering is an indication of which datapoints group together based on similarities between their feature vectors. If successful, it allows us to interpret groupings of datapoints by integrating results with other classes of data. For examples, a cluster of proteins with a majority of them being involved in a specific biological process may lead us to conclude that the remaining unannotated proteins are candidates to also play a role in that process (Jain and Dubes, 1988).
Among a wide variety of clustering algorithms hierarchical clustering is often a method of choice as it does not require assumptions about the number of clusters. Also, having set a cut-off point we are able to explore hierarchical structure below that cut-off (within the cluster) down to the level of a single datapoint (Jain, 2010).
Furthermore, we can use the knowledge of relationships between features for insight which will allow us to tune parameters of the analysis for the outcome to have higher domain-specific validity. The principle of hierarchical clustering can be generalised to any arbitrary metric function. The metric function can be learned from data, if we have any indication of correct response values, or can be derived based on prior insight about data Johnson (1967). Finally, observations based on clustering outcomes and prior knowledge of research goals may lead to use simplified metrics to describe datapoints, thus defining more easily interpretable measures than through typical dimensionality reduction procedures such as Principal Component Analysis or Multidimensional Scaling.

3.1.2 *Postsynaptic density*

Since the research presented here is focused on global proteome-wide effects as well as synaptic function evolution it is suitable to use an established classification to identify proteins present in the synapse which will allow to compare synaptic proteins to the remaining human proteome members. As much as there is no consensus regarding which proteins can be reliably associated with this anatomical grouping, here I use the group of proteins present in human PSD identified in Bayés et al. (2012) experiments. The main advantage of this list is the source tissue coming from human, additionally, it is fully experimental, not computationally inferred (examples of alternative lists not used here: Yoshimura et al., 2004; Collins et al., 2006).

3.1.3 *Ontologies*

In order to structure domain knowledge we can formalise it in a form of an ontology (Chandrasekaran et al., 1999). Mathematically, a biological ontology can be described as a directed acyclic graph of ontology terms where edges represent relationships between terms such as *'is a member of'*. Each term node has a set of biological entities (proteins/molecules) associated to it.
Entry to node associations can be derived automatically, e.g. through NLP approach and a large dataset of publications but also through manual curation using expert knowledge. In the biological setting Gene Ontology (GO) was the first big project in this field motivated by increasing availability of experimental data and two use cases: sharing discovered associations in a structured way, and using these associations for interpreting results of further studies (Ashburner et al., 2000). Many other biological ontologies followed; in this chapter I am using a selection of them, and the choice is motivated by contents of the graph and the quality of annotation.

- **GO** (Blake et al., 2015) is the primary target of bioinformatics exploratory analyses due to its breadth, depth and generality of terms; consists of three graphs for 3 classes of terms - Biological Process, Molecular Function, and Cellular Component; here I use the Biological Process ontology. Multiple slimmed down versions exist (terms reduced to the domain of interest), hence here **niGO** (Geifman et al., 2010) will be of special interest - it is limited to immunological and neural terms and term selection is curated manually by the authors, it contains 4935 terms to be tested for enrichment compared with 26837 in full GO for human, mouse and rat (Geifman et al., 2010).

- **Reactome** (Joshi-Tope et al., 2005; Mi et al., 2017) is an ontology where terms represent molecular pathways, i.e. proteins and other chemical compounds belonging to one compound are linked together through actual molecular interactions within a shared pathway.

- **Panther** (Thomas et al., 2003) is based on multiple levels of distinguishing proteins into classes and subclasses based on their origin and shared functional domains.

- **Human Disease Ontology (HDO)** (Osborne et al., 2009) aggregates information about gene-disease associations from literature, terms are diseases and groups of diseases with common aetiology or symptoms.

Although it is not an ontology, KEGG pathway encyclopaedia will be used in this chapter in a similar way (Kanehisa and Goto, 2000; Kanehisa et al., 2017) supplementing Reactome pathway-focused analysis.

3.1.4 *Enrichment analysis*

The most common use case for ontology users is interpreting a grouping of genes or proteins, asking a question whether these genes over-represent a specific annotation class, e.g. a biological process in GO or a pathway in Reactome. Membership might be fuzzy where instead of imposing a cutoff of a given value we attach values to all genes. In this case an established procedure is to use Kolmogorov Smirnov test to determine whether the distribution of scores for genes in the term is significantly different from the background distribution - this approach is called Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). In a simplified scenario of discrete membership in gene sets we can test over-representation of terms through one of the common testing schemes for contingency tables such as Fischer exact test where probability of deviation from the null hypothesis (the same representation of a term in the sample set as in the background) is calculated exactly. Both methods are implemented in common software packages such as `topGO` (Alexa et al., 2006).

3.1.4.1 *Classic vs elimination*

When using the established enrichment testing procedure described above two major possibilities exist for matching members of the input set to the nodes in the ontology tree as we traverse it. In the classic method all associations of a given node and its descendants are used to compute test statistic. In the elimination method the ontology graph is traversed in a bottom up direction. At any stage all nodes at one level of the graph can be processed at the same time as they are not connected. If a node is found significant, all genes associated with it are removed from the annotation of its ancestors. Once all nodes in one level are tested the procedure repeats on the following level (Alexa et al., 2006). Figure 3.1 illustrates an example from authors of the method. The method reduces the number of less specific high-level nodes with significant enrichment, instead focussing attention on the most specific terms which are enriched.

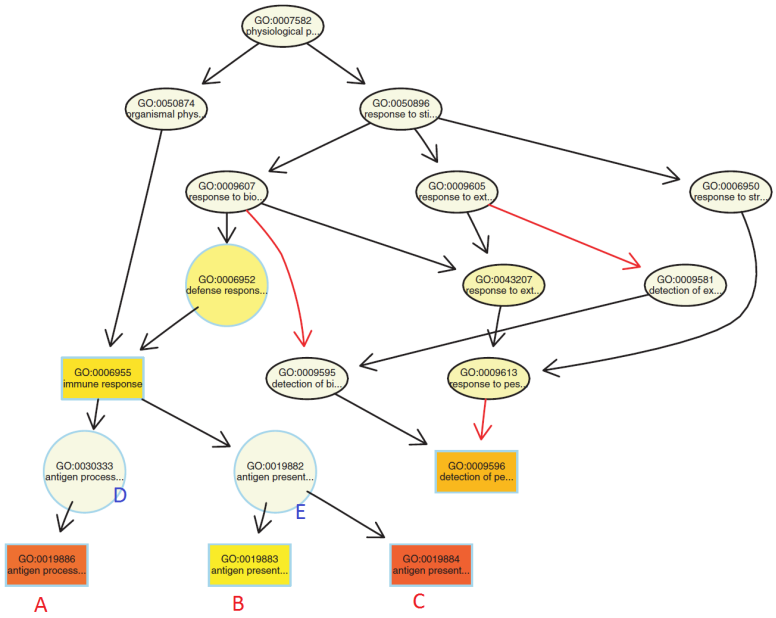


Figure 3.1: Eliminaltion method of enrichment testing in ontologies. Bottom nodes (A, B, C) are found to be significant so genes associated with them are removed from the list of associations on the following level, as a result these nodes (D, E) are not significantly enriched. From Alexa et al. (2006)

3.1.5 *Multiple testing corrections*

Due to the exploratory nature of GSEA the results normally need to be corrected for multiple comparisons. However, it is unclear how multiple corrections should be addressed, especially when using an elimination method which is indirectly guarding against spurious effects. The two main approaches are Holm-Bonferroni method and False Discovery Rate (FDR) with Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The former can be summarised as following: all p-values are ordered: $p_1, ..., p_n$, for the significance level α we set k as the minimal index for which $p_k > \frac{\alpha}{n+1-k}$, null hypotheses are rejected for $i = 1, ..., k-1$, whereas they are accepted for $i = k, ..., n$. It is generally considered a conservative method, yet not as conservative as Bonferroni correction which it is based on (it can be seen as a stepwise application of Bonferroni correction). FDR is defined as the expectation of a proportion of false discoveries among the discoveries. If we set it at α then in th Benjamini–Hochberg procedure we order p-values similar as before, then we find the largest k such that $p_k < \frac{k\alpha}{n}$, null hypotheses are rejected (i.e. discoveries are made) for $i = 1, ..., k$. FDR-adjusted values of p-values (often called q-values) are cal-

culated according to an improved method introduced by Yekutieli and Benjamini (1999) which ensured monotonicity of $q(p)$ function.

3.1.6 *Objectives*

This chapter follows exploratory research inquiry into temporal patterns of selection pressure across the evolutionary history through aggregation of results of aBSREL modelling. This is achieved through two main objectives:

1. The first goal is to use unsupervised methods in search of broad patterns and clear groupings of datapoints - evolutionary profiles of proteins.
   The steps to achieve this goal include the choice of a clustering method and a distance metric motivated not only by their performance but also by computational considerations given the dataset size. Following that, the objective is to justify a selection of cluster number, again, multiple factors need to be considered. Clustering will provide protein groupings which we will use in the subsequent parts of this chapter - for enrichments and for detection of differences which characterise PSP.

2. The second goal is to derive informative features from the temporal feature vector which summarise biologically valid aspects of episodic selection pressure, then test how we can use it to describe the timeline of the human proteome (and the synaptic proteome) diversification.
   Further to that, these features will allow to identify proteins of interest based on extreme values of the measures.

Then, enrichment analysis will help to interpret data-driven groupings of proteins based on clustering and timeline features. Enrichment testing will indirectly contribute to the overarching question of conservation of functional complexes of proteins and their co-selection.

As the thesis is primarily concerned with synaptic function, the auxiliary objective is to establish whether post-synaptic proteins can be distinguished from the background using either differential representation of clustering groups or any other fea-

tures based on temporal modelling data.

3.2 RESULTS

3.2.1 *Data - episodic selection pressure model*

Since the overarching research question of the thesis is concerned with the evolution of synaptic function in human, I am only concerned with evidence for episodic selection pressure on the path from the root of the phylogenetic tree to the human leaf (see Figure 3.3), however, using these results it would be possible to study different endpoints (discussed in section 6.3.4).

First, outputs of aBSREL modelling are mapped back to the reference tree of life. It is possible to do this as all input trees for modelling were subtrees of the full tree of 69 taxa created by pruning from species for which there was no annotated orthologs of a given protein (see section 2.2.5). The only ambiguity remains in cases of a missing taxon which is the sole leaf of the tree responsible for a specific divergence point.
For example, if for a given protein there was an ortholog available for both *T. strychita* and *P. anubis* but not *C. jaccus*, the input tree for aBSREL had only one branch between divergence points from the two former taxa. As a result modelling results map from that single branch map back to both sub-branches in the reference tree as there is no way to determine which of the two branches is responsible for episodic selection signal in absence of data for *C. jaccus*.

A feature vector for each protein consists of 22 values of log-likelihood ratio for consecutive branches of the tree from root to human, the magnitude of these values represents support for episodic positive selection on each of these branches.

3.2.1.1 *Protein origin measure*

On top of the feature vector, each protein can also be described by a single discrete measure informative of the earliest point on the path from root to human in the tree (see Figure 3.3) where I can make an assumption that an ortholog of a given protein was present. This point will be denoted by the distance (number of branches) from human to the node which is the most recent common ancestor of human and the

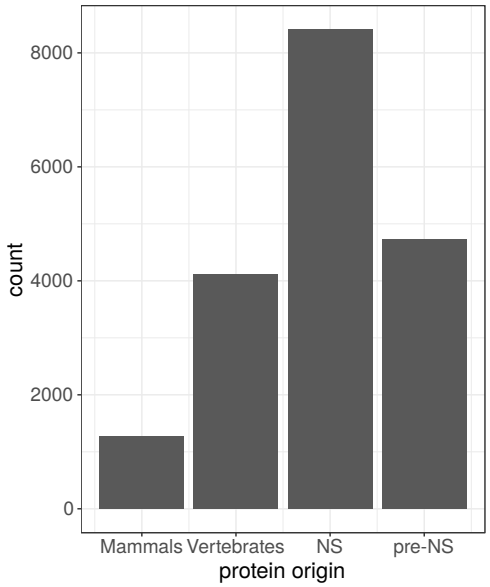
Table 3.1: Taxa mapping to divergence points on the human root path. The number of the common ancestor node is the distance from the *H. sapiens* leaf of the tree along the path to its root (1-indexed). Also see Supplementary Table A.3 for full names of all taxa used.

Origin node	Taxa diverging at the node
1	*H. sapiens*
2	*P. troglodytes*
3	*G. gorilla*
4	*P. abelii*
5	*N. leucogenys*
6	*M. mulatta, P. anubis, C. sabaeus*
7	*C. jacchus*
8	*T. syrichta*
9	*O. garnettii, M. murinus*
10	*T. belangeri*
11	*D. ordii, O. princeps, O. cuniculus, C. porcellus, R. norvegicus, I. tridecemlineatus, M. musculus*
12	*S. araneus, E. europaeus, V. pacos, S. scrofa, P. vampyrus, M. lucifugus, O. aries, T. truncatus, F. catus, E. caballus, B. taurus, A. melanoleuca, M. furo, C. familiaris*
13	*C. hoffmanni, E. telfairi, P. capensis, D. novemcinctus, L. africana*
14	*M. eugenii, S. harrisii, M. domestica*
15	*O. anatinus*
16	*A. carolinensis, A. platyrhynchos, M. gallopavo, P. sinensis, T. guttata, F. albicollis, G. gallus*
17	*X. tropicalis*
18	*L. chalumnae*
19	*G. morhua, A. mexicanus, O. latipes, D. rerio, T. nigroviridis, X. maculatus, P. formosa, T. rubripes, O. niloticus, L. oculatus, G. aculeatus*
20	*P. marinus*
21	*C. savignyi, C. intestinalis*
22	*C. elegans, D. melanogaster*
23	*S. cerevisiae*

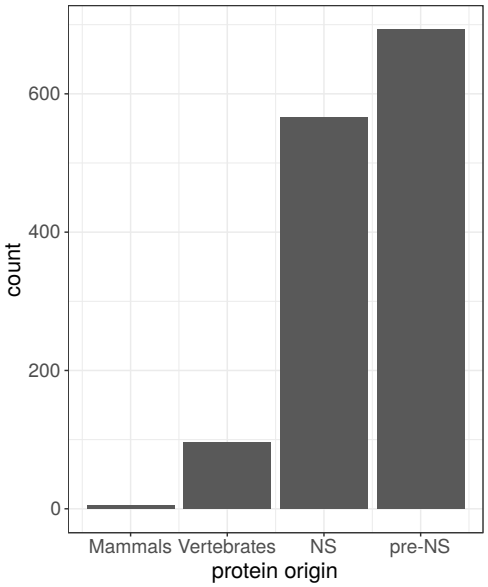
furthest taxon with an ortholog of a protein present. Table 3.1 lists mappings of origin points of proteins to sets of taxa which diverged from the linear root - human path at a given node (nodes are numbered starting from *Homo sapiens* - node number 1) Most important breaking points in that linear scale are the beginning of the mammalian clade, the beginning of the vertebrate clade, and the beginning of organisms with a nervous system. Therefore throughout this thesis I often bin proteins' origin by these breaking points:

- Mammals for $1 \leqslant \mathtt{orgin} \leqslant 15$
- Vertebrates for $16 \leqslant \mathtt{orgin} \leqslant 20$
- Organisms with nervous system (denoted as NS in tables and figures) for $21 \leqslant \mathtt{orgin} \leqslant 22$
- Organisms prior to nervous system development (denoted as pre-NS in tables and figures) $\mathtt{orgin} = 23$.

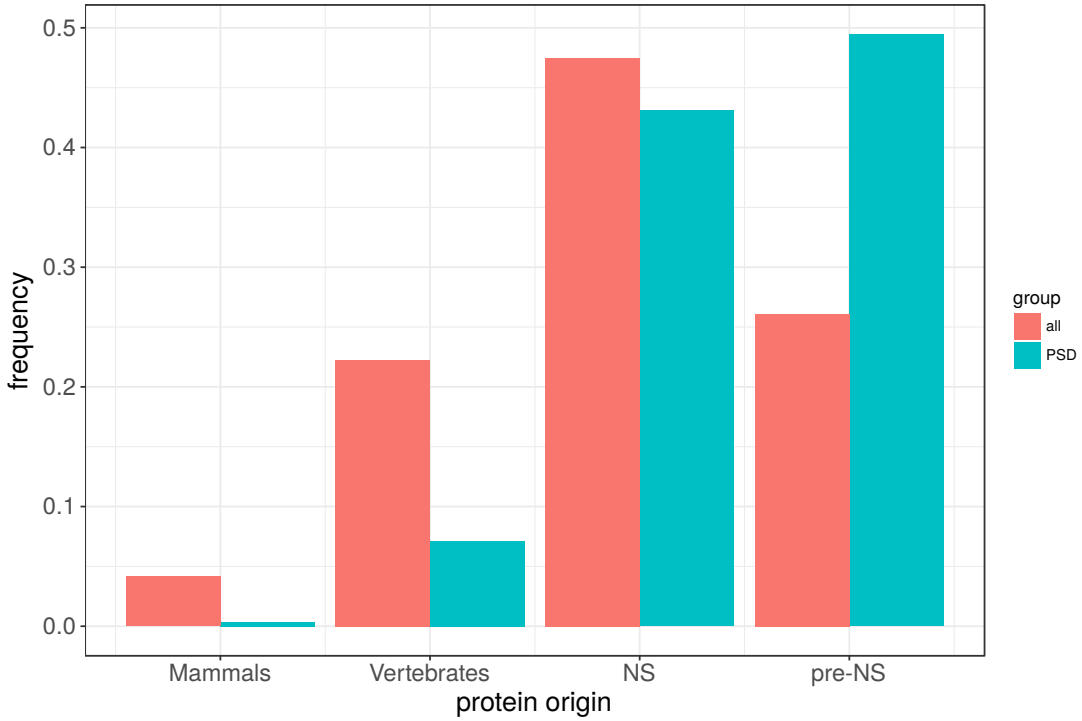
When comparing the full proteome to the PSP there is a striking difference, as synaptic proteins have a higher proportion of proteins originated in the earliest two categories compared to the full proteome, and similarly much lower proportion of recently originated proteins (See Figure 3.2). Based on the origin measurement alone, synaptic function as a whole appears to be conserved deeper than others.



(a) Counts in the full proteome



(b) Counts in the PSP



(c) Comparison with both distributions of origin normalised to frequencies

Figure 3.2: Distribution of origin of all human proteins (a), only postsynaptic proteins (b); (c) shows counts of the full proteome from (a) and the PSP from (b) converted to relative frequencies and plotted on the same axes for contrast. See section 3.2.1.1 for explanation of origin categories on x-axis in these plots.

3.2.1.2 *Significance filtering*

According to Smith et al. (2015) the log-likelihood ratio statistic which indicates the strength of evidence for the positive selection model is distributed according to a mixture of chi-square distributions, thus, I could assign a numerical p-value to each log-likelihood ratio and filter values lower than the corresponding test statistic threshold. This procedure revealed 3242 proteins with no significant evidence for diversification on any of the branches since their emergence until *H. sapiens* - for all branches LRT statistic was below the significance threshold. However, I did not treat them as negatively selected throughout, instead I interpreted these cases as lack of conclusive evidence and therefore excluded from many analyses (such as clustering). Finally, for most branches which fell under the significance threshold of detecting evidence for positive selection, LRT statistic was still greater than 0; where this was the case I set it to 0 for all further analyses according to the equation below (lrt is log-likelihood ratio test statistic i.e. $-2\ln(likelihoodratio)$)).

$$lrt_{new} = \begin{cases} 0, & lrt_{raw} \leqslant threshold \\ lrt_{raw}, & \text{otherwise} \end{cases}$$

3.2.2 *Clustering*

The next part of this project aimed at identifying genes experiencing a similar pattern of their timeline of diversification according to the results of aBSREL branch-by-branch analysis. In a novel approach to this research question values of LRT statistic of the selection model vs. the neutral model for branches on the root-human path (see Figure 3.3 and Table 3.2) were used for classification as a feature vector of length 22. After enforcing a significance threshold to log-likelihood values as mentioned in section 3.2.1.2 I used 14,394 proteins for clustering analysis (81.6% of all proteins with temporal data available).

3.2.2.1 *Distance measure*

Hierarchical clustering depends on the distance metric employed, as optimising distance metric is not a central point of this work, cosine distance measure was selected

Table 3.2: Mapping of nodes of the human path to the timeline of divergence (from Ensembl Compara based on median estimate from available literature). Multiple nodes with the same divergence time and same taxonomic name (such as rows 2 & 3, or 10 & 11 in this table) arise from resolution of ambiguous trichotomies in the tree. Also, compare Figure 3.3

Node distance	Scientific name	Ensembl name	Divergence time (mya)
1	*Homo sapiens*	Human	0.0
2	Homininae	Hominines	8.8
3	Homininae	Hominines	8.8
4	Hominidae	Great Apes	15.7
5	Hominoidea	Apes	20.4
6	Catarrhini	Apes&OW mokeys	29.0
7	Similiformes	Simians	42.6
8	Haplorrhini	Dry-nosed primates	65.2
9	Primates	Primates	74.0
10	Euarchontoglires	Primates&Rodents	92.3
11	Euarchontoglires	Primates&Rodents	92.3
12	Boreoeutheria	Placental mammals	100.0
13	Eutheria	Placental mammals	104.2
14	Theria	Marsupials&Placentals	162.6
15	Mammalia	Mammals	167.4
16	Amniota	Amniotes	296.0
17	Tetrapoda	Tetrapods	371.0
18	Sarcopterygli	Lobe-finned fish	414.9
19	Euteleostomi	Bony vertebrates	441.0
20	Vertebrata	Vertebrates	535.7
21	Chordata	Chordates	722.5
22	Bilateria	Bilateral animals	937.5

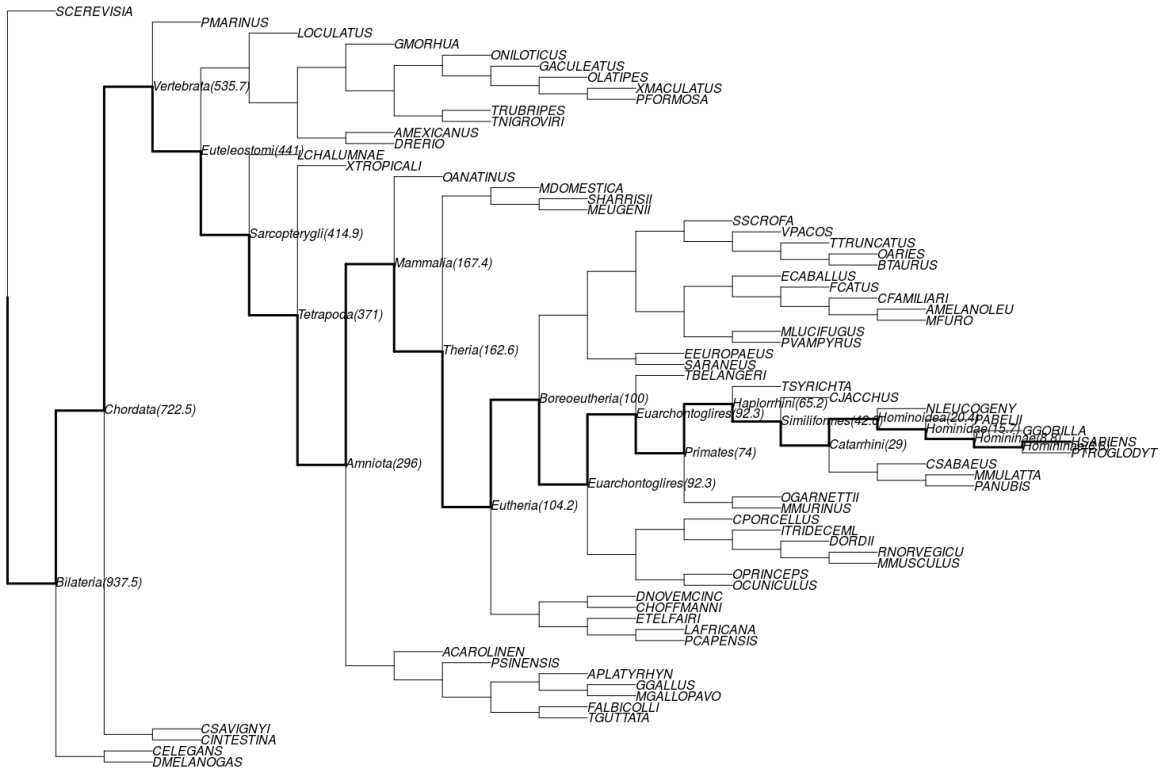


Figure 3.3: Path from root to human according to Ensembl Compara; highlighted branches form a path from the root to the *H. sapiens* leaf. Also compare Table 3.2 and see Supplementary Table A.3 for full names of all taxa used.

for capturing reasonably high dimensionality (Aggarwal et al., 2001), and lesser focus on outlier values of one particular feature; since feature vectors consisted of log-likelihood ratios this was an important consideration. Cosine distance is defined as *1-cosine similarity* where *cosine similarity* is defined as following:

$$\mathrm{similarity}(X_1, X_2) = \cos(\theta) = \frac{X_1 \cdot X_2}{\|X_1\|_2 \|X_2\|_2}$$

Another practical advantage of cosine distance is that, similar to correlation coefficient, values of cosine similarity are strictly bounded, in the general case the range between -1 and 1, yet for positive vectors (such as data I use here) only between 0 and 1. By extension the cosine distance defined as above is also bounded to $[0, 1]$ interval.

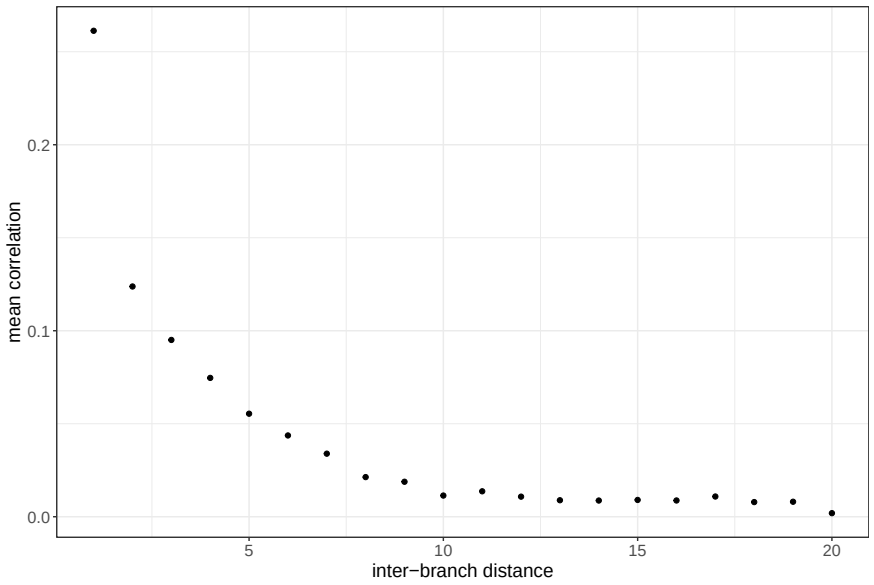


Figure 3.4: Interbranch correlations on the linear path from root to human. Correlations between 14394-long feature vectors, one for each branch, averaged by the distance between branches on the path from root to human. Relatively high correlation can be observed for close branches, it decreases gradually with the inter-branch distance.

3.2.2.2 *Temporal aspect of features*

Cosine distance is a good starting point when working with relatively high dimensional feature vectors, however, in this case I wanted to utilise additional information from the temporal meaning of the feature vector, i.e. similarity between $i-th$ feature for one protein and $i+1-th$ feature for another protein is less indicative of shared selection pressure timeline than similarity between same $i-th$ features. Figure 3.4 illustrates validity of this claim as branches close to each on the root-human path other exhibit correlation in episodic positive selection.

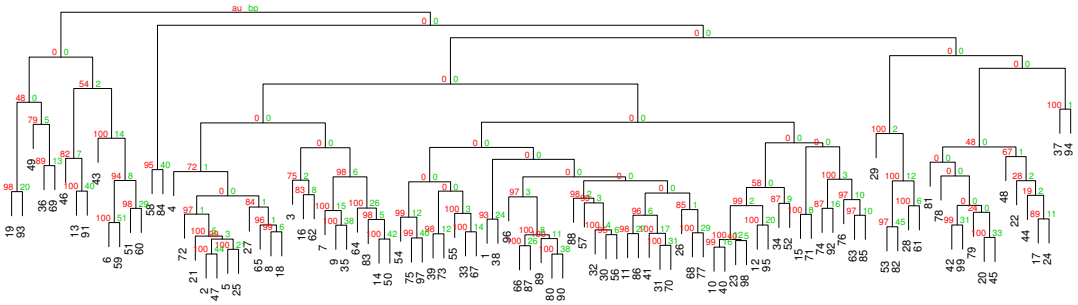
After experimenting with cross-correlations and edit distances, which are supposed to serve the purpose of capturing sliding pattern matches, I found a simple solution of smoothing the vector with a discrete Gaussian kernel of length 5 to be sufficient in accounting for temporal similarities, efficient in execution, and elegant in interpretation. In summary, the distance matrix was a 1:1 weighted sum of a cosine distance measured between all raw pairs of temporal selection feature vectors and a cosine distance measured between all smoothed temporal feature vectors. All values in the matrix were bounded to $[0, 2]$ interval.

3.2.2.3 *Hierarchical clustering*

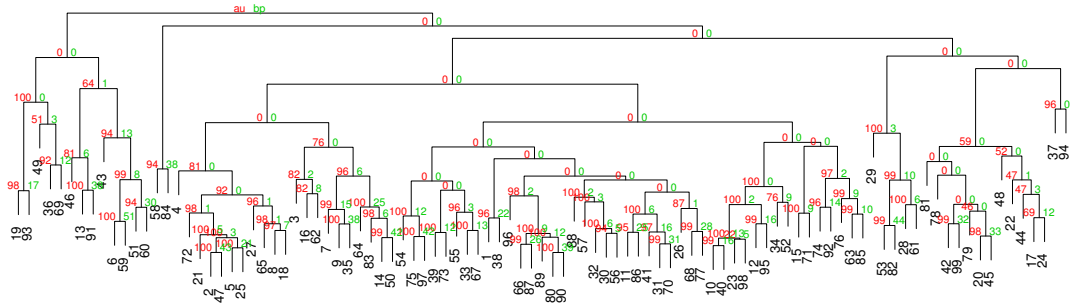
Perhaps the most commonly used hierarchical clustering algorithm is the Ward method (Ward, 1963; Murtagh and Legendre, 2014) which is an agglomerative clustering procedure. Using the error sum of squares objective function, the distance matrix is assumed to contain Euclidean distances between datapoints which is not the case here. Although there are published attempts to adjust Ward's method to other metrics and successfully apply it to real data problems (Strauss and Von Maltitz, 2017) to the best of my knowledge there is no consensus whether it can extend to any arbitrary distance metric. Another alternative solution which would allow use of Ward method is to transform the original distance matrix to a Euclidean one as implemented by Dray and Dufour (2007) in `ade4` package.

For sake of simplicity and to avoid additional steps in the procedure, instead of using Ward method I opted for a more robust linkage method which does not put any constraints on the distance matrix.

Two alternatives of average linkage and complete linkage are available, (Johnson, 1967; Murtagh, 1983) here, I tested both methods, and in the context of cluster number choice (discussed in the following section) I selected the average linkage method (compare Tables 3.3 and A.2) .



(a) $N_{bootstraps} = 250$



(b) $N_{bootstraps} = 500$

Figure 3.5: Bootstrap clustering tests for a sample of 100 proteins. Trees for further values of $N_{bootstraps}$ are available in the Supplementary Figure A.1, comparison with a larger sample of proteins (200) available in Supplementary Figure A.2

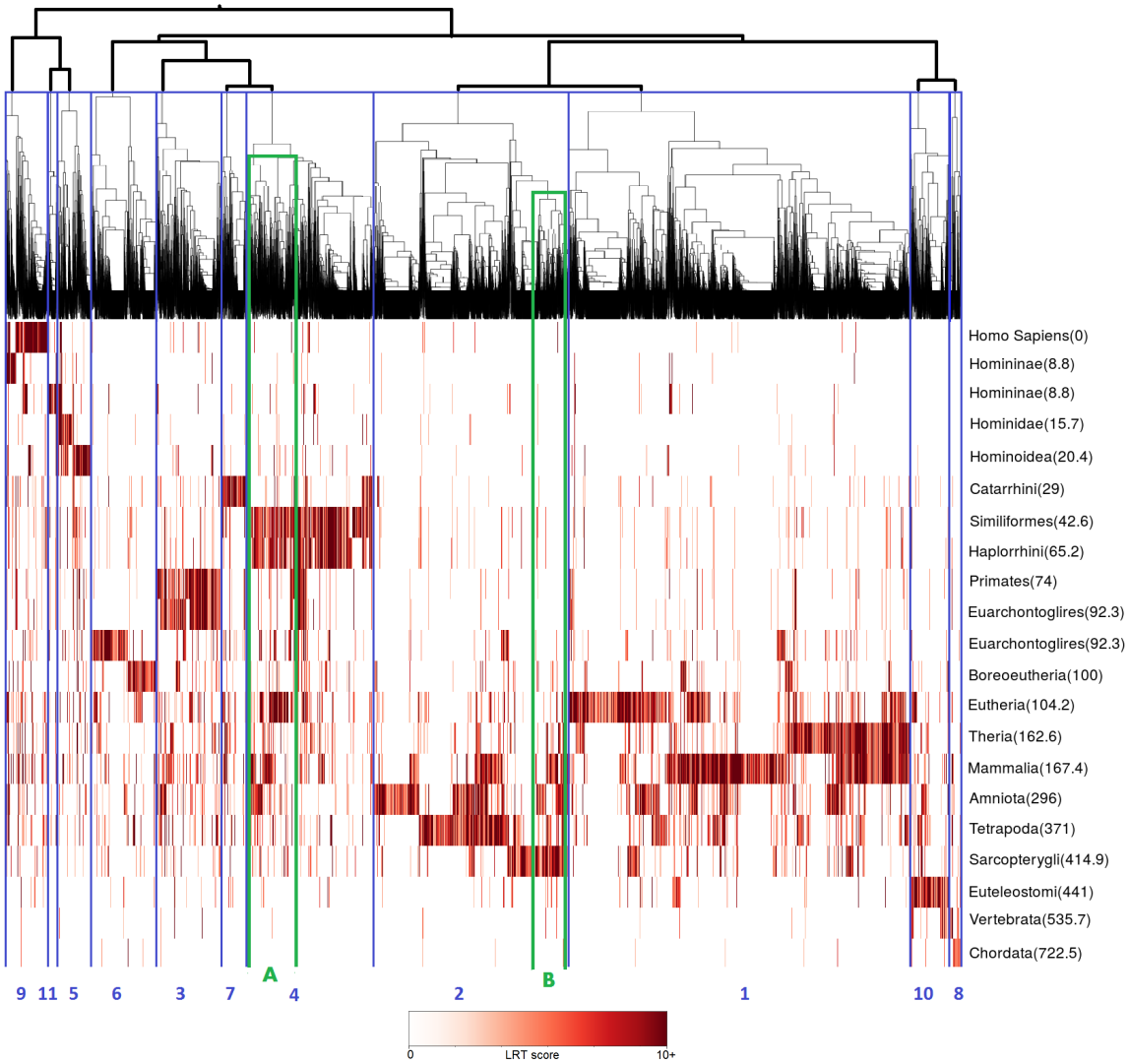


Figure 3.6: Clusters of temporal diversification profiles according to average linkage hierarchical clustering and a cut at the height of 11 clusters. Each column in the heatmap represents a protein and subsequent rows represent evidence for positive selection on consecutive branches of the root to human path. Numbers in blue are cluster numbers referenced in the text. Letters in green point to clusters at much lower cut point which demonstrate subjectively assessed bimodality.

3.2.2.4 *Cluster number selection*

The initial idea for determining the number of clusters was to use multiscale bootstrapping (Shimodaira, 2004; Suzuki and Shimodaira, 2006). The advantage of the method is that it is a fully data-driven approach, where clusters which remain consistent across bootstrap sample clusterings are selected (normally using 95% or 99% confidence criterion). Figure 3.5 contains an example of bootstrapped clustering for a small (100) subset of proteins for $N_{bootstraps} = 250$ and $N_{bootstraps} = 500$. Sup-

plementary Figures A.1 and A.2 contain further clustering trees for this sample of proteins as well as another larger sample (200 proteins).
We can observe how the confidence criterion metric for intermediate nodes of the clustering tree generally increases with the number of bootstraps. For the smallest number of bootstraps only very low level nodes reach acceptable confidence level. For the larger set of proteins (Supplementary Figure A.2) the intermediate nodes acquire increased confidence level slower than in the smaller set, when matched for the number of bootstraps. This illustrates it was not realistic to perform the same procedure for the entire dataset. Not only does the calculation of the distance matrix scales with the size of the matrix but also the number of required bootstrap samples grows to the point where it is not feasible to run such analysis for the full set. An alternative of executing it for a sub-sample of the full human protein would force adoption of a clustering method which assigns remaining datapoints to sample-derived clusters (Jain, 2010) which in itself is hard to reconcile with the hierarchical clustering paradigm.

Instead I opted for choosing the cluster number according to a mixture of data-driven and interpretative criteria. The first criterion was keeping the number of clusters minimal yet at the same time keeping balance between sizes of clusters. In hierarchical clustering very consistent small clusters tend to break off very early while traversing the tree from top to bottom thus creating a situation with a few clear small clusters and the remainder cluster which is harder to partition thus creating gross inbalance in cluster sizes. The Ward method partly guards against this situation at the same time compromising intra-cluster cohesion, however, as discussed in the previous section, here I use one of the linkage methods due to the chosen distance metric.

Transition between 10 and 11 clusters in the average linkage method marks a drop in maximum of all maximum intra-cluster distances, a drop in maximum median of intra-cluster distances, and also splits the largest cluster (see Table 3.3). No other K above $K = 11$ satisfies these three criteria at once. $K = 14$ would also be a viable choice (mainly because of the split of the largest cluster), yet in cases such as this, simplicity of a model can be used as an additional criterion, hence a solution with fewer clusters is preferred. A similar table for the complete linkage method is presented in the Appendix (Supplementary Table A.2), interestingly, for the same number of clusters (11) the complete linkage solution achieves a worse maximum of maximum

Table 3.3: Selection of the cluster number for the average linkage method. K is the number of clusters in the tree cut, then measures of intra-cluster distance follow, max(max) refers to maximum of all maximum distances in each of the K clusters. min(median) is the lowest out of all median distances in each of the K clusters, respectively max is the highest one. min(n_i) is the size of the smallest cluster, and respectively max(n_i) is the size of the largest one. Compare Table A.2 for similar statistics for a different method of hierarchical clustering.

K	Intra distance: max(max)	min(median)	max(median)	min(n_i)	max(n_i)
5	2	0.75	1.28	658	8384
6	2	0.75	1.074	658	8384
7	2	0.65	1.034	658	8384
8	2	0.414	0.986	159	8384
9	2	0.414	0.986	159	8384
10	2	0.414	0.986	159	8384
11	1.929	0.414	0.833	159	5341
12	1.929	0.414	0.833	159	5341
13	1.929	0.414	0.755	159	5341
14	1.923	0.414	0.75	159	3795
15	1.923	0.414	0.741	159	3795
16	1.923	0.36	0.741	53	3795
17	1.923	0.36	0.741	2	3795
18	1.923	0.36	0.707	2	3795
19	1.923	0.36	0.707	2	3795
20	1.923	0.36	0.707	2	3795

distances and maximum of median distances. Its only advantage would have been more compact and balanced size of clusters which however comes at the price of their cohesion, therefore I decided for average linkage solution with $K = 11$.

3.2.2.5 *Cluster profiles*

The clustering, visualised in Figure 3.6, abstracted the dominant episodic diversification period for proteins. It also succeeded in exploiting the temporal ordering of features, with high values of temporally adjacent features putting two proteins in one cluster. Cluster 2 is a good example where highly significant episodic diversification in either of neighbouring branches 16, 17, or 18 puts a protein in the same cluster even though there are relatively few proteins with episodic selection in more than one of these branches. Also, considering higher cuts (higher meaning fewer splits) of the hierarchical tree clusters 9, 11, and 5 fall in the same supercluster, collectively representing all proteins with dominant period of diversification across the 5 most recent branches.

At this level of cutting the tree, no clear clusters with multimodal patterns (assessed subjectively), representing multiple periods of diversification separated by periods of negative pressure, were extracted. However, when exploring the hierarchical clustering tree deeper (lower cuts) we can observe these multimodal patterns to emerge, e.g. leftmost egde of cluster 4 (green rectangle A in Figure 3.6) and rightmost edge of cluster 2 (green rectangle B in Figure 3.6) get separated from their main cluster because of their bimodality. I will explore these features of diversification further when designing useful summary metrics of selection pressure profiles in section 3.2.4.

3.2.3 *Ordering and grouping*

Clusters were successful at identifying the dominant period of diversification for a protein, and in case of smaller clusters for lower cuts of the hierarchical tree also multiple dominant periods of diversification. Using these clusters as a guidance proteins (or protein groups) can be ordered by their dominant period of positive episodic selection pressure. However, such ordering becomes ambiguous when a protein displays multiple separate diversification periods so I aimed to further exploit the temporal ordering of consecutive elements of the feature vector to impose more meaningful

ordering between proteins. Here I used a measure of most recent positive diversification for that purpose.

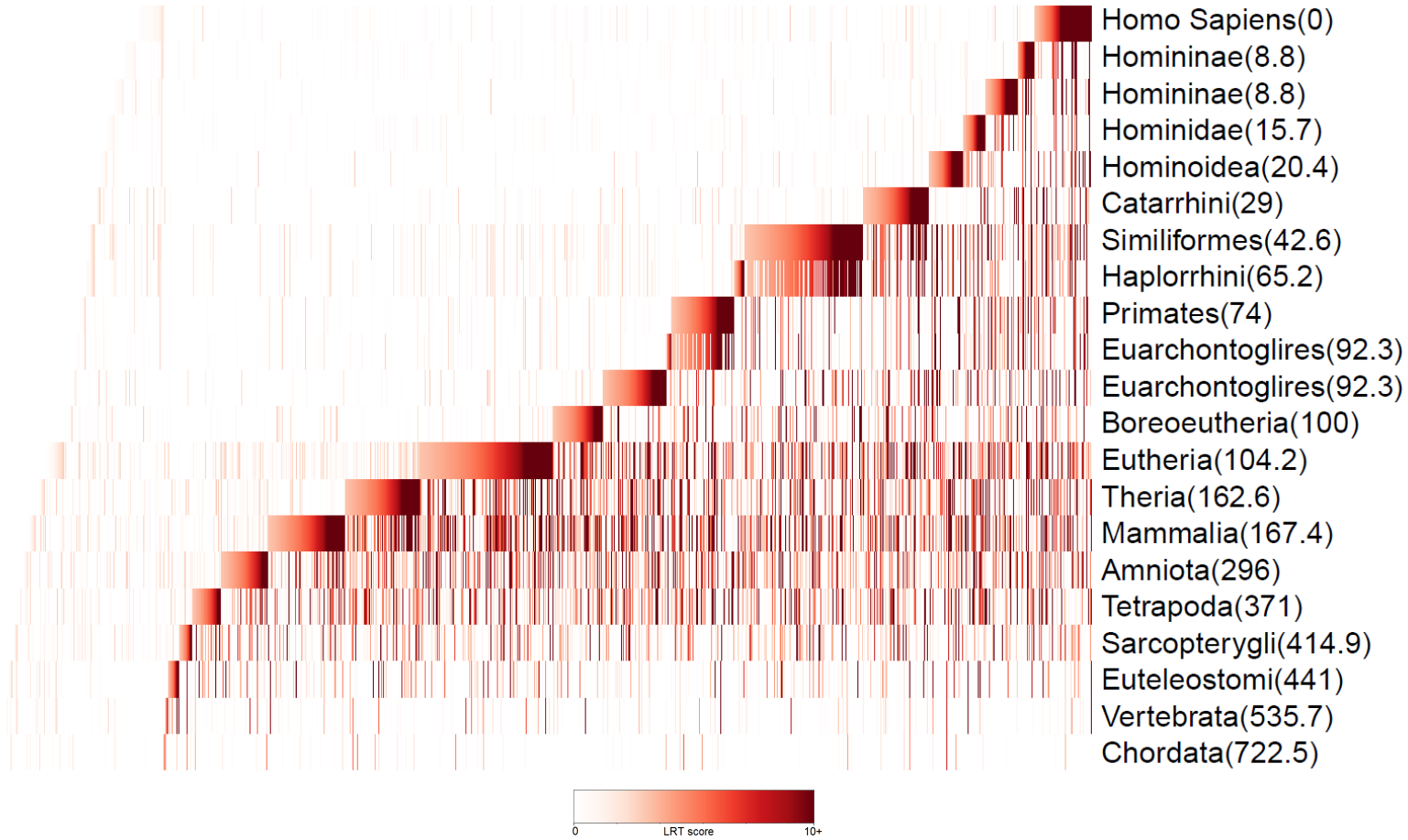
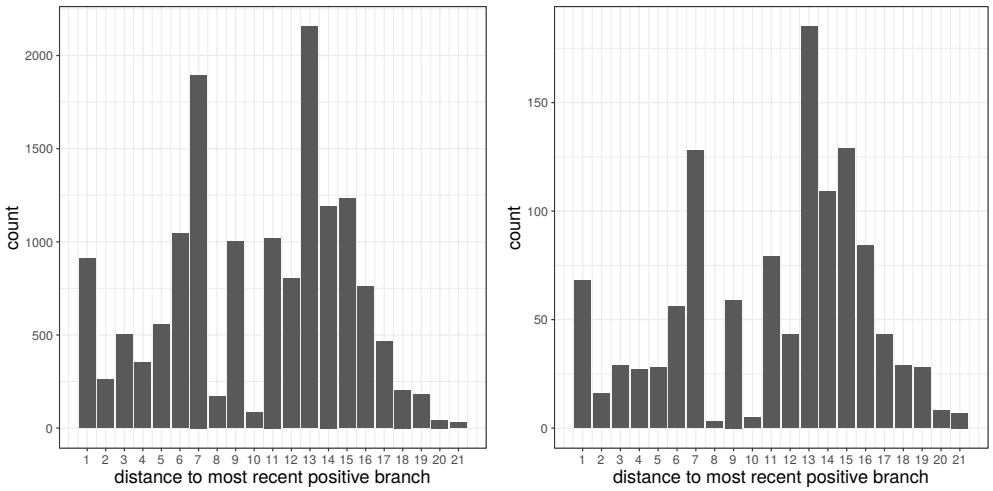
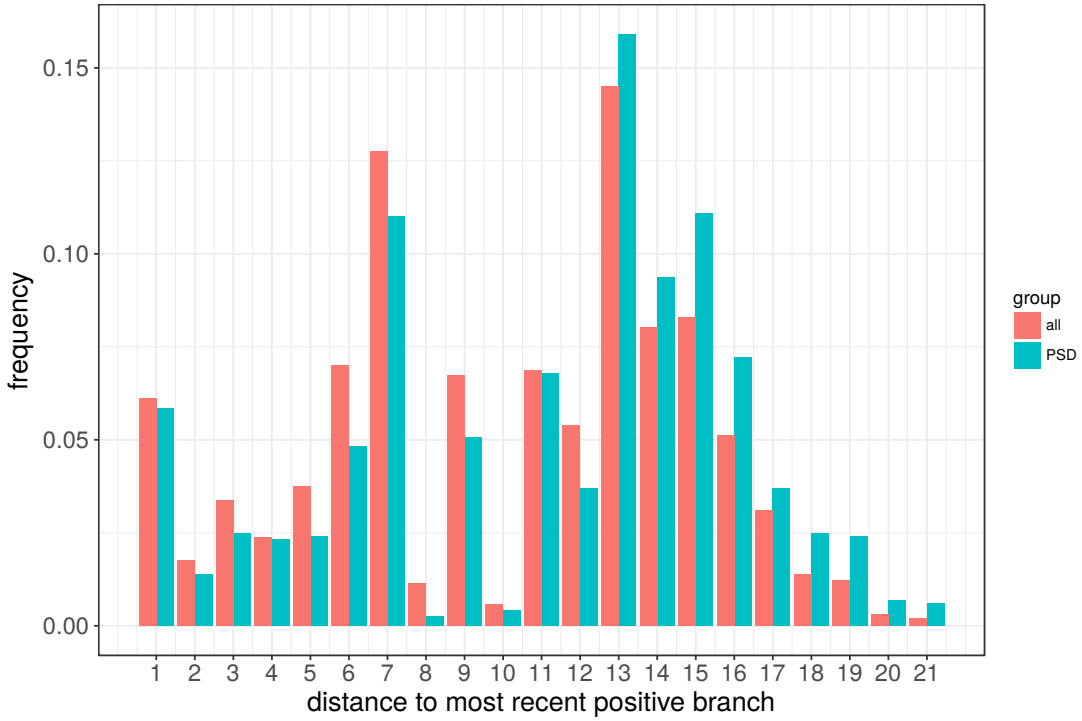


Figure 3.7: Most recent positive diversification for the full proteome. X axis - proteins ordered by MRP, Y axis - consecutive branches of the root-human path in the phylogenetic tree, see figure 3.3 and table 3.2. for further explanation.
Intensity of colour represents strength of evidence in favour of positive episodic selection on that branch (log-likelihood ratio). Non-significant values of log-likelihood were not included in ordering but were deliberately left in the plot (hence a large group of proteins on the left), see section 3.2.1 for more detailed explanation of the variables used for plotting and section 3.2.3 for rationale for ordering of the data.

Most Recent Positive Selection/Most recent branch with evidence for positive selection (MRP) measure describes the distance from the *H. sapiens* leaf to the most recent branch on the path from root to human for which I inferred significant positive selection pressure. The *H. sapiens* node is indexed with 1 therefore $\mathrm{MRP} = 1$ represents the state where the most recent positively selected branch is the one leading up to *H. sapiens* leaf.



(a) Counts in the full proteome (b) Counts in the PSP



(c) Comparison with both distributions normalised to frequencies

Figure 3.8: Distribution of MRP in all human proteins (a), only in the PSP (b), and in (c) counts of the full proteome and the PSP from (a) and (b) respectively are normalised to relative frequencies and plotted together on the same axes; note the peaks of the distribution at the same values of MRP for both sets. Also see section 3.2.3 for explanation of the measure used here.

Figure 3.7 visualises the result of ordering datapoints by most recent positive diversification with each column representing a temporally ordered feature vector for that protein. The secondary ordering variable is the magnitude of evidence for positive selection pressure on the most recent significantly positive branch. Proteins without any branches with significant episodic selection pressure (which were excluded from clustering analysis) occupy the leftmost section of the heatmap.
Highest density of evidence for positive selection is observed between Tetrapoda and Eutheria branches, a lot of proteins do not have significant positive branches beyond this point, another step change occurs around the Similiformes branch. For the following primate branches there are generally few proteins completing their positive diversification up until the *H. sapiens* branch where there are relatively many proteins under positive selection. However, the difference is exaggerated visually because many proteins positive e.g. in the Hominoidea branch have more recent evidence for positive selection too, hence the strip of proteins completing their diversification at that branch (for which $\mathrm{MRP} = 5$) appears thin.

In order to better distinguish between all evidence for diversification and MRP in Figure 3.8 I compare the distribution of MRP between the full proteome and PSP in a similar way to Figure 3.2. Here, qualitatively the PSP distribution is the same as for the entire proteome with a small trend for slightly higher frequencies of the PSP compared to the full proteome for earlier MRP and the inverse being true for more recent MRP (see Figure 3.8c), which I hypothesise is the confounding effect of stark difference in origin distribution between the full proteome and the PSP (the dependency is further discussed in section 3.2.3.1 and visualised in Figure 3.9). Both in full proteome and in PSP there are two evident peaks, at $\mathrm{MRP} = 7$, and $\mathrm{MRP} = 13$, which are also visible as long stretches of proteins for rows *Simiformes* and *Eutheria* in Figure 3.7. Both peaks fall on biologically relevant divergence points - the recent one on a transition between Simians (New World and Old World monkeys) and other more more distant primates such as tarsiers and lemurs (prosimians), and the early one one a transition between Marsupials and Placental mammals. Especially the latter was a major milestone in animal evolution which is of general interest, however, the former is equally interesting from the point of view of the nervous system and cognitive-behavioural changes.

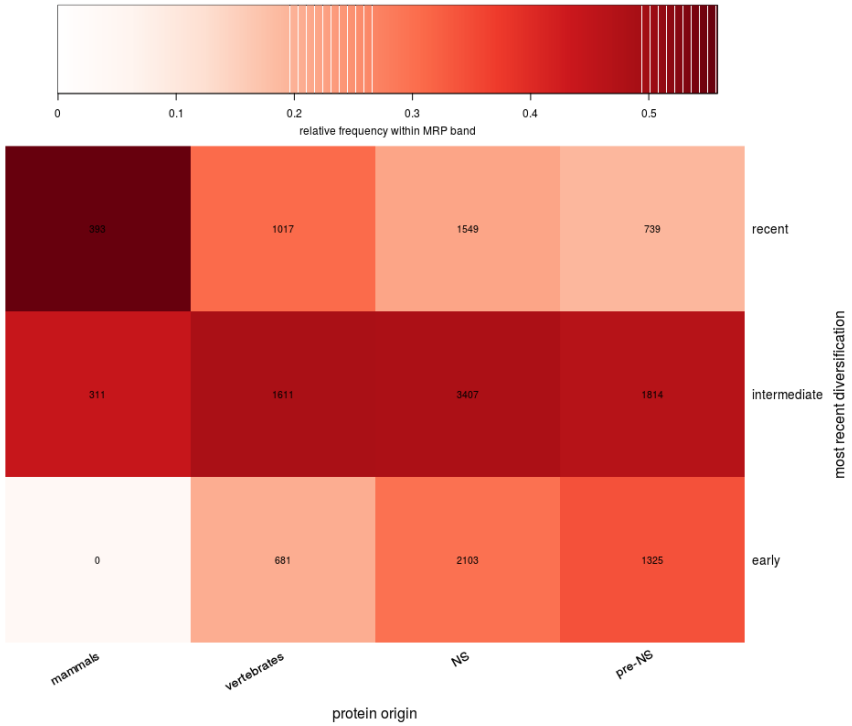


Figure 3.9: Protein origin against evidence for the most recent positive diversification (MRP) binned into biological breaking point categories for origin and data-driven bins for MRP; displayed numbers are counts contributing to the specific cell of the heatmap, colours represent magnitude of relative frequency in each cell after normalising counts over columns.

3.2.3.1 *Origin and most recent diversification dependency*

Similarly to protein origin (compare section 3.2.1.1 and Figure 3.2), I binned MRP into broader categories, however, unlike for protein origin, bin boundaries are derived from the distribution of MRP not from biological meaning of specific nodes. The procedure resulted in dividing the range of MRP into 3 bins:

- Recent diversification for MRP $\leqslant 6$ i.e. after the most recent peak at MRP $= 7$
- Intermediate diversification for $7 \leqslant$ MRP $\geqslant 13$ i.e. including 2 main peaks at MRP $= 7$ and MRP $= 13$
- Early diversification for MRP $\geqslant 14$ i.e. prior to the earlier peak at MRP $= 13$

In the heatmap in Figure 3.9 I compare binned protein origin and binned most recent positive selection. The effect of a dependency between the most recent diversification and origin point is clear but not unexpected. Proteins of deep origin complete their diversification earlier in absolute terms. There also seems to be a pattern where

generally there is a certain period following the protein's appearance where it is under positive selection before it becomes conserved under purifying selection. This observation motivated deriving a measure of the temporal aspect of diversification which corrects for a bias associated with the origin point of a protein.

3.2.4 *Measures*

On top of already mentioned MRP other measures were also extracted for each of the studied proteins to help interpretation of temporal results. I was particularly interested in the overall amount of time a protein was under diversifying selection since its appearance. It can be quantified in a simple way by counting branches under significant selection. However, proteins originated at different points through the evolutionary history. Thus it is also a valid question to ask whether there is a period following protein origin during which proteins are under diversifying selection until they become conserved, and how the length of this period can differentiate proteins.

Also, if selection pressure on a molecular level reflects specific events at the environmental and phenotypic level then it is relevant to think about bursts of diversifying pressure. It might take a long time for the active diversification to complete on a molecular level, hence significant selection on a few adjacent branches but possibly only the peak of the period of postitive selection is relevant, and their presence can be conceptualised in similar ways to the positive branches - most recent one, total, etc.
The following measures allowed me to quantify the subjective observations about the timeline of diversification of a given protein and summarise them in a single numerical value.

- **Total number of positives** - the total number of branches with significant evidence for positive episodic selection pressure.
- **Diversification window** - related to the total number of positives but it only describes the difference between the distance to origin of the protein and the distance to the most recent positive branch, i.e. $\mathtt{Origin} - \mathtt{MRP}$.
- **Total number of peaks** - Peaks were identified as local maxima in a vector representing data series. First, the data series vector was differentiated with a discrete approximation of the first derivative (differences between adjacent

values); then, the sign function was applied, and the resulting vector was differentiated again in the same way. Peaks were identified as positions of values in the final vector equal to -2 (i.e. places where sign switched from positive to negative). In case of a series of two or more adjacent positions in the output where original data series values were at the same level only the first one of a series is returned.

- **Most recent peak** - the most recent out of peaks identified according to the principle described above.

The first two measures were then normalised by the value of protein origin to remove dependency on how many branches on the root-human path were available. This way I gained two more measures:

- **Relative total number of positives** which is equal to the total number of positives divided by the distance to protein origin, i.e. $\frac{\mathrm{TotalPositives}}{\mathrm{Origin}}$.
- **Relative diversification window**, similar to the above, it is equal to the length of th diversification window divided by the distance to protein origin, i.e. $\frac{\mathrm{Origin}-\mathrm{MRP}}{\mathrm{Origin}}$

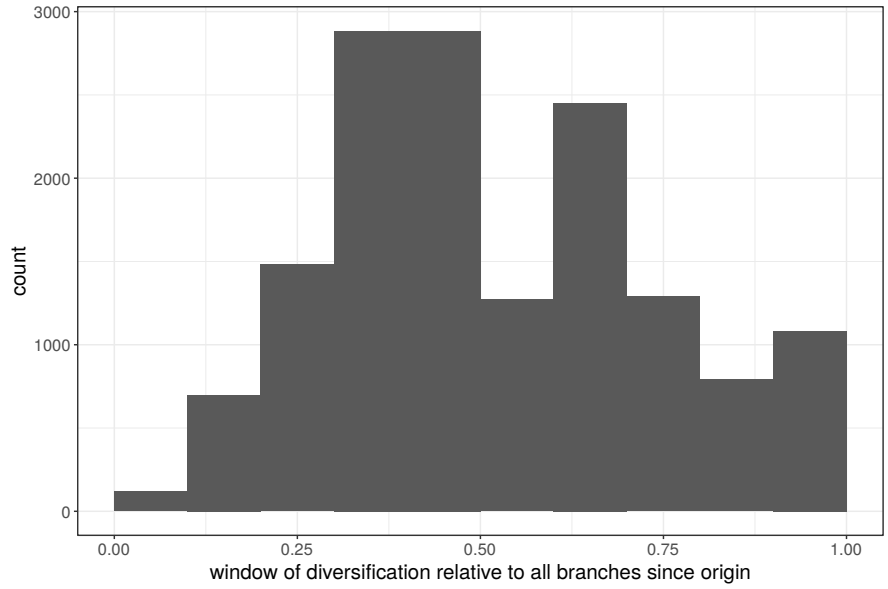
Distributions of the last two measures over the entire human proteome are visualised in Figure 3.10, respective distributions for PSP are not qualitatively different.

These measures will be used throughout the thesis to describe diversification profile of certain proteins and its relationship to other properties of proteins. They will be treated in a non-parametric way, not making assumptions about their distribution fits, hence I will use median and Inter-quartile Range (IQR) to summarise them. However, it is worth noting that the relative total number of positives closely fits gamma distribution (see Supplementary Figure A.3).

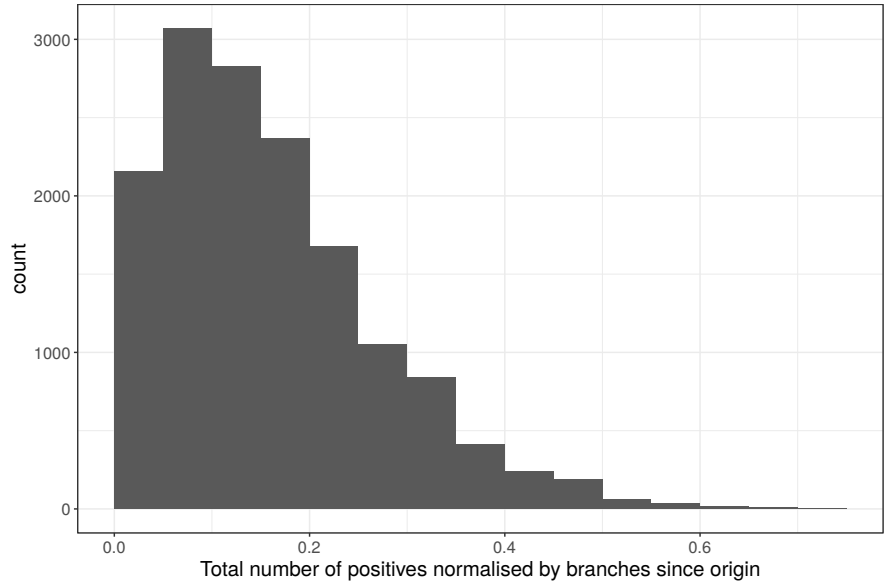
3.2.4.1 *Interesting genes based on extreme values of measures*

Proteins which could be interesting because of combinations of extreme values of timeline measures were identified, I distinguished four classes of proteins, the first two being generally recently positively selected and the remaining two generally experiencing very early positive selection.

1. **Diversifying throughout** - The first group contains proteins with a long relative diversification window and a high value of relative total number of positive



(a) *Relative diversification window*



(b) *Relative total number of positives*

Figure 3.10: Distribution of timeline measures defined in section 3.2.4 - *Relative diversification window* in (a) with relatively irregular shape and indication of bimodality, and *Relative total number of positives* in (b) distributed approximately according to gamma distribution (also see Supplementary Figure A.3).

branches (within that diversification window). The proteins are extracted based on condition that at least one of these measures needs to be in the 10th decile, and both need to be in at least the 9th decile for respective measures. It represents proteins which diversified a lot throughout their lifetime, corrected for the point at which they appeared. Deciles are adjusted arbitrarily to keep the output sets small enough for ease of interpretation of the results.

2. **Diversifying only recently** - The second group contains proteins with a long diversification window (early origin, recent MRP) but with only one (recent) *peak* of positive selection. It represents early origin proteins which were under purifying selection up until a recent burst of episodic positive diversification. I use a criterion based on peak as opposed to number of diversifying branches to allow for longer 'runs' of positive branches.

3. **Conserved throughout** - In the following group I put proteins which represent opposite features to the first group, i.e. both the relative diversification window and relative total number of positive branches is low, at most the 1st decile for one of them but at most 2nd decile for the other.

4. **Only early diversification** In the final group, which is conceptually similar to the previous group (conserved throughout) I did not use relative measures, only restricted the period of positive diversification to maximum 2 branches following the origin of a protein. This group can also be considered a representation of the opposite extreme pattern to the second group (diversifying only recently).

Members of the PSP which satisfy criteria for these four groups are listed in table 3.4 together with counts for respective sets in full human proteome.

3.2.4.2 *Clusters and measures*

In section 3.2.2.5 I described an unsupervised clustering procedure as an extraction tool for the most dominant period of diversification for a given protein, which is often composed of a few consecutive branches with evidence of positive diversification.

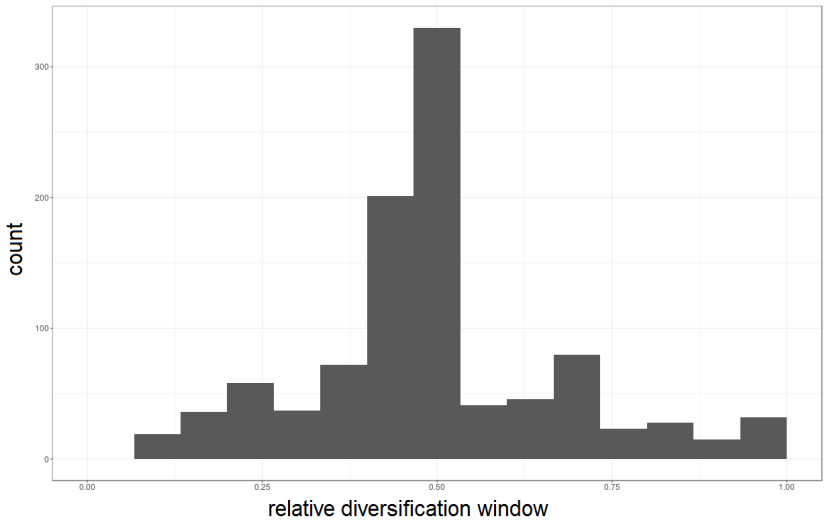
Proteins within each cluster can represent a wide array of origin points. Here I compare distributions of measures correcting for the origin of a protein introduced in section 3.2.4 for different clusters to determine if a purely unsupervised method using a base feature vector captures some of the more complex effects discussed when constructing timeline measures. Indeed, there are substantial qualitative differences between distributions of relative diversification window, see Figure 3.11. Particularly cluster 6 displays a clear bimodal distribution of the variable suggesting that despite being the smallest cluster, it captures two distinct subgroups of genes. The main feature of this cluster in Figure 3.6 is a dominant very early burst of episodic selection pressure - proteins for which that was the most recent diversification are responsible for the first mode in Figure3.11a, however, the second mode is comprised of genes

Table 3.4: Groups of interesting proteins in the PSP based on extreme values of the timeline measures, see section for group inclusion criteria.

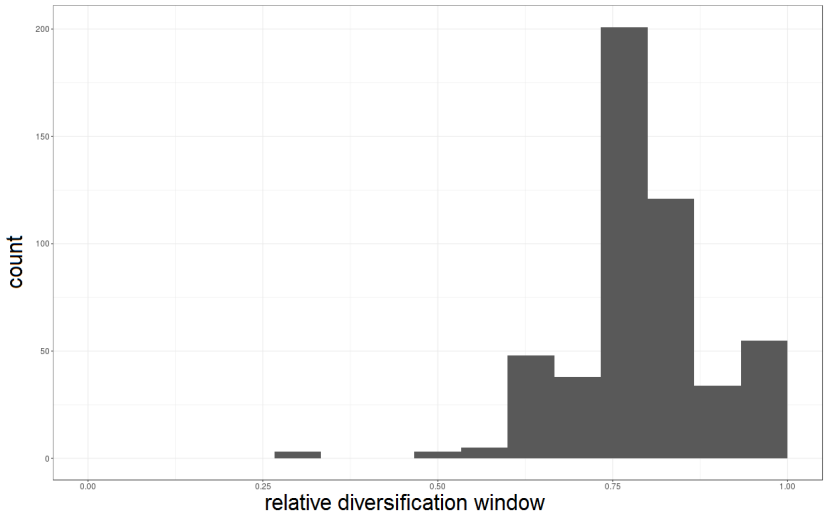
Group	Count (all)	Count (PSP)	PSP members
Diversifying throughout	429	22	ANKRD24, AP3D1, TMEM245, EIF2AK2, DNAJB2, EPB42, AMER2, FMNL1, FGG, LIMCH1, KIAA1217, NCAM1, MPRIP, PALM, PLCD3, PPFIA4, SIRT2, SORBS1, SGIP1, SPTB, TAOK2, TJP2
Diversifying only recently	266	8	CRMP1, DOCK1, EPO, KIAA1598, NDUFA9, RAB5C, STX1B, WDR7
Conserved throughout	409	28	ACTR3, ARF5, CAMK2A, CAPZA2, DLD, DOCK3, GNAI2, HSPA9, KPNA1, KIF5C, MYO1B, NSFL1C, OLA1, PHB, PRMT5, RAB11B, RAB14, PPP3CA, SEPT11, RPL38, RPL10A, RPS13, TAOK1, UBC, TUBB4B, YWHAH, VSNL1, WDR48
Only early diversification	240	16	ACTN1, ACTN4, ARF5, ATP1B2, APPL2, CEND1, CTNNB1, DDRGK1, DTNA, IRGQ, SEPT11, RPL10A, RPS13, SLC1A3, TUBB4B, WIPF2

which also diversified further on. Also, comparing clusters 5 and 2 which are both relatively large clusters, apart from a shift in the mean between them cluster 5 has much lower variance than cluster 2 which means that not only does it capture genes with dominant episodic diversification around nodes $6, 7, 8$ but also most of these proteins originated at a similar point in the tree.

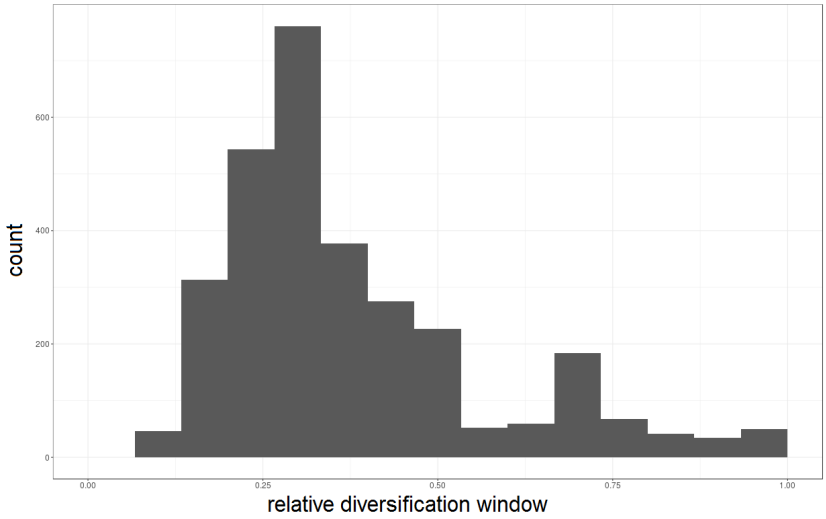
Table 3.5 summarises numerical properties of clusters as measured on intra-cluster distance metric 3.2.2.1 as well as measures introduced in section 3.2.4. Note, there are large difference in non-parametric summary statistics for relative diversification window and a relative total positives distributions.



(a) Cluster 6



(b) Cluster 5



(c) Cluster 2

Figure 3.11: Distributions of *Relative diversification window* for different clusters, see Figure 3.6 for general cluster profiles and section 3.2.2 for an explanation of the clustering procedure.

Table 3.5: Statistics of timeline clusters

	Members		PSP representation	Distance		Diversification window		Total positives		Pre-NS proteins	
Cluster	Full	PSP	corrected p-value	median	max	median	IQR	median	IQR	number	fraction
1	5341	421	0.9096	0.755	1.929	0.409	0.152	0.136	0.113	1370	0.257
2	3043	285	0.0108	0.741	1.914	0.318	0.194	0.1	0.132	941	0.309
3	1008	61	0.0754	0.65	1.885	0.591	0.062	0.19	0.132	232	0.23
4	1991	99	0	0.707	1.923	0.682	0.049	0.217	0.17	412	0.207
5	527	23	0.0108	0.833	1.903	0.783	0.076	0.182	0.199	97	0.184
6	1026	61	0.0624	0.826	1.897	0.478	0.141	0.105	0.119	190	0.185
7	389	28	0.8523	0.595	1.713	0.727	0.055	0.15	0.174	62	0.159
8	165	36	0	0.672	1.756	0.182	0.273	0.087	0.091	95	0.576
9	658	54	0.8523	0.75	1.844	0.952	0.031	0.19	0.136	152	0.231
10	627	90	0	0.65	1.841	0.348	0.281	0.13	0.095	291	0.464
11	159	10	0.7841	0.414	1.681	0.864	0.026	0.14	0.176	30	0.189

3.2.5 *Ontology Enrichment*

In order to test the hypotheses about the relationship between functional grouping of proteins and their common diversification timeline I performed extensive enrichment testing with multiple ontologies (see section 3.1.4) using cluster groupings as well as levels of MRP as a grouping factor.

3.2.5.1 *Enrichment methodology*

I executed separate analysis for full proteome and PSP, where for PSP I used only PSP members as background. I used the `topOnto R` package for enrichment tests and ontology graph manipulation (He and Simpson, 2017b,a). It closely follows the approach of `topGO` (Alexa et al., 2006) but allows for arbitrary ontology structure use. In all analyses I used the elimination method and corrected p-values of enrichment tests for multiple comparisons using the FDR paradigm.

niGO I reduced the full GO graph by picking specific manually curated nodes related to neural and immune functions (Geifman et al., 2010). Reducing the ontology graph to terms which are relevant for the use case, here, neural function, serves two purposes. First, it helps further interpretation and ensures that relevant terms are not suppressed by the irrelevant ones. From the statistical point of view, fewer terms tested translate to less strict multiple testing correction with no loss of information if we were not interested in the remaining terms to begin with.

HDO (reduced) There was no readily available procedure for reducing HDO therefore here I opted for a simple rule which exploits the fact that ontologies can be easily traversed as graphs, I simply picked all descendants of two nodes: *Nervous system disease* and *Cognitive disorder* which together add up to 1136 terms (see Supplementary Figure A.4 for Venn diagram of the overlap).

Panther In case of Panther (Thomas et al., 2003), the protein class ontology, I used the full ontology graph.

Pathways: KEGG & Reactome I used the full ontology graph for Reactome (Joshi-Tope et al., 2005; Mi et al., 2017). KEGG (Kanehisa and Goto, 2000; Kanehisa et al.,

2017) does not have an ontology graph structure, it is just a set of terms, still, they can be tested for enrichments as if all terms were mutually disconnected and at the same level in an ontology.

3.2.5.2 *Enrichment results summary*

Below I list most interesting terms at the top positions of enrichment lists for clusters, all uncorrected p-values were significant at $p < 0.05$ level, but here FDR-adjusted p-values are presented for reference (which in most cases are non-significant). The choice of reported clusters, MRP thresholds, and terms is subjective and guided by the goals of the thesis as presentation of the full results would not be informative. Cluster numbers refer to the annotation in Figure 3.6.

niGO Due to the high number of terms even in the reduced graph FDR adjustments were particularly harsh for this ontology. Full proteome tests revealed enrichment of *cognition* and *B cell receptor signaling pathway* as top terms according to the elimination method in cluster 9 - the most recent one (fdr=1 for both). The top enriched term in another recent cluster, number 5, was *G-protein coupled receptor signaling pathway* with relatively low fdr=0.23.

reduced HDO Test for PSP revealed that the top enriched term for Cluster 8 was *autosomal recessive juvenile Parkinson disease* (fdr=0.39). Another noteworthy result with relatively low FDR was appearance of *frontal lobe epilepsy* as the top term for MRP $\leqslant 2$ (fdr=0.31). There was no clear pattern for widely studied neurodegenerative diseases such as Parkinson's or Alzheimer's disease - apart from *autosomal recessive juvenile Parkinson disease*. Their ontology terms never surfaced at the top of the list yet they appear relatively high in lists for multiple clusters and MRP values.

Panther Tests based on the full proteome revealed the following top enriched protein classes of interest:

- Cluster 8 - *ATP synthase* (fdr=0.025) and *actin and actin related protein* (fdr=0.075)
- Cluster 9 - *Cadherin* (fdr=0.52)
- Cluster 1 - *G-protein coupled receptor* (fdr=0.15)

Additionally, the test limited to PSP revealed the following terms at the top of the ranking for Cluster 8 - *ribosomal protein* (fdr=0.06), *microtubule family cytoskeletal protein* (fdr=0.62), as well as more general *cytoskeletal protein* (fdr=0.62).

Pathways: KEGG & Reactome The only clear result for Reactome was the enrichment of a large number of terms associated with translation (such as *Eukaryotic Translation Termination*, *Ribosomal scanning and start codon recognition*, *Translation initiation complex formation* and others) in clusters 10 and 8, as well as for MRP $\geqslant$ 20. FDR-adjusted values reached significance levels in these protein sets which was not the case for other clusters and MRP thresholds. KEGG provided similar results yet with less granularity, *Ribosome* pathway was significantly enriched for MRP $\geqslant$ 19.

Overall, the outcome of enrichment testing was overwhelmingly inconclusive for all tested ontologies and for all tested groups of genes, possibly with the only exception for early diversifying groups (clusters 8, 10, and high MRP values). This was mainly due to very strong multiple testing correction imposed due to large sizes of ontology graphs, however, even if not statistically significant, the trend which emerged will guide interpretation of results in this chapter.

3.3 DISCUSSION AND CONCLUSIONS

Research presented in this chapter was the first stage of exploration of the modelling data generated according to the procedure described in Chapter 2.
I used episodic selection evidence data, and applied unsupervised methods in the form of hierarchical clustering to group genes into distinct profiles of episodic diversification. The ontology enrichment test application did not reveal a substantial link between shared temporal diversification profile and common protein function, class, or involvement in disease based on multiple biological ontologies.
Also, I extracted biologically motivated measures from episodic selection feature vectors which allowed me to identify general trends in the diversification timeline for both the full proteome, and PSP, as well as helped me compare the two sets. Finally, synaptic proteins characterised by extreme profiles of episodic selection based on these measures were selected for further investigation.

3.3.1 *Early diversifying functions*

The most consistent observation across enrichment analyses was enrichment of basic cellular functions (such as ribosomal pathways) among early diversifying groups, both when proteins were divided based on their MRP, and when they were grouped based on hierarchical clustering. Their function is fundamental to the basic function shared among different types of cells in the organism hence they appear early and complete diversification under positive selection early; past the shift towards vertebrates there was not much scope for them to diversify any further.

3.3.2 *Peaks of most recent diversification*

It is interesting to see many proteins diversifying deep into very recent divergence points. Perhaps contrary to expectations, the anatomically selected set of post-synaptic proteins followed the same peaks of most recent diversification, even though representation of clusters from section 3.2.2 was different compared to the full human proteome (see Table 3.5) and the distribution of protein origin differed substantially (see Figure 3.2). This points to a universal quality of these two peaks. Indeed the earlier peak coincides with the end of an evolutionary transition which diverged marsupials from placental mammals. The second, more recent peak is interesting for two reasons.

First, we can interpret its significance in the context of species involved at the affected divergence points as it falls at the border between tarsiers and lemurs (collectively described as prosimians) and simians (ie. New World monkeys and Old World monkeys). According to studies based on anatomical properties of recovered fossil remains the early prosimians were largely tree-dwelling nocturnal animals with limited social behaviour (foraged alone), contemporary descendants include lemurs and tarsiers which largely confirm these observations (Müller and Thalmann, 2000). However, the early simians were almost exclusively diurnal, and gradually developed complex social behaviour (Ross, 1996), there is also evidence for social foraging in present day catarrhines as well as their early counterparts (Ross, 1996). This explains selection pressure put on their sensory systems - different visual system (colour vision) and reduced dependence on olfaction (confirmed in a study of receptor range by Rouquier et al., 2000). Also, selection pressure would have acted on molecular

mechanisms of complex cognition such as learning from their peers in social groups.

Second, dating of the divergence point which marks the ends of branch number 7 points to an incredibly interesting period in Earth's geological history, known as Cretaceous–Paleogene (K–Pg) extinction which is dated around 65-66 million years ago (Schulte et al., 2010; Renne et al., 2013). Median dating of nodes 7 and 8 which limit branch 7 is 42.6mya and 65.2mya which is a relatively wide time interval yet it encompasses the period immediately following the mass extinction. Changes in Earth climate put different demand on all systems of living organisms and could encourage diversifying selection in proteins across the entire proteome. Opening of niches previously occupied by extinct species would have a similar effect on molecular evolution of the proteome. Four previous mass extinctions (end Ordovician, late Devonian, end Permian, and end Triassic) are much more difficult to pin to specific branches of the tree due to a very sparse temporal resolution of deeper divergence points of the tree of life. However, I hypothesise that with better resolution of divergence points and by extending analysis to all divergence points across the tree outside of the root-human path, branches including periods immediately following each of the mass extinctions would show similar enrichment to the one discussed here. On the other hand we only have access to reference transcriptome/proteome data for contemporary species which means they all survived (or diversified after) all extinction events thus the effect on MRP measure would be the strongest for the most recent extinction.

3.3.3 *Identifying interesting proteins based on evolutionary profiling*

In the analysis of extreme temporal profiling (see Table 3.4) the primary suspects of interest are proteins with abundant evidence for positive episodic selection - group one (diversifying throughout). However, if a protein has been conserved and under purifying selection for a long time then it is a protein that is very sensitive to any deviation and any smallest mutation in its upstream regulatory process may lead to pathology - despite the focus on detection of positive selection, proteins with highly limited evidence for it are also interesting. In this context proteins in the second group ('diversifying only recently') are especially interesting as they remained under purifying pressure until very recently when there was an episodic diversification event which can be attributed to an explosion in cognitive and behavioural function

in primates (Bradshaw and Rogers, 1993).

Starting with the two groups characterised by dominant purifying selection (3 and 4), *RPL/S* proteins appear in both groups. They are ribosomal proteins which play part in the most fundamental function of the cell - production of proteins (Wool et al., 1995; Wool, 1996). It is complementary evidence to the enrichment analysis outcomes. Furthermore, in the same two groups there a numerous proteins associated with the cytoskeleton - actin and microtubule structures abundant on both sides of the synaptic density: First, in the 'conserved throughout' group there is ACTR3 - a major part of the ARP2/3 complex involved in cytoskeleton growth (Mullins et al., 1998); then in the last group ('only early diversification') there are *ACTN1* and *ACTN2* - F-actin cross-linking proteins (Huang et al., 1997). Finally, *TUBB* appears in both of these groups - as a major constituent of microtubules it affects vesicle transport and mitochondria transport (Morris and Hollenbeck, 1995).
The overall conclusion based on interpreting PSP members of conserved groups is that many of the proteins responsible for maintenance of the basic cellular machinery and cell shape did not experience much diversification or at least not past the initial period immediately following their appearance.

Then, among 'diversifying throughout' proteins (group 1) we can spot the widely studied *NCAM1* protein which is a stimulator of tyrosine kinase activity involved in neurite outgrowth, and has been associated with schizophrenia (Sullivan et al., 2007), as well as alcohol and nicotine dependence (Yang et al., 2008a). Abundant evidence for diversifying selection throughout the tree on this protein is particularly interesting from the clinical perspective and contributes to the discussion of psychiatric disorders arising as a by-product of increasing complexity of neural function.
Finally, the second group ('diversifying only recently') offers an interesting selection of synaptic proteins. First, there is *CRMP1* which mediates reelin signalling in cortical neuronal migration (Yamashita et al., 2006). Su et al. (2007) showed that mice depletion of *CRMP1* impaired their long-term potentiation on the molecular level and impaired spatial learning and memory on the behavioural level. This points to the importance of circuit setup in achieving complex cognitive function. Also, there are three proteins involved in exocytosis in this group - *STX1B*, *WDR7*, *RAB5C* (De Camilli and Jahn, 1990). Although in this context they are members of the PSP, they are also pre-synaptic proteins. Recent positive selection following a long period of

purifying selection observed in the exocytosis process may suggest that although the regulatory role of these three proteins remained conserved up until recently, recent environmental and behavioural constraints in primate species more closely related to human put new demands on the regulation of this process which was reflected as positive pressure on the molecular level.

3.3.4 *Methodological limitations*

Enrichment analysis methods such as contingency table tests and GSEA are often criticised as inadequate methodology for their typical use cases (Goeman and Bühlmann, 2007; Tamayo et al., 2016). The structure of terms is inherently difficult to integrate within statistical testing framework. As a result, tests traditionally designed for datasets with far less complex dependency in them might suffer from lack of power, which is further aggravated by multiple testing corrections in a scenario where technically their independence assumptions are violated.
For ontologies n is the number of all nodes of the ontology graph for which the enrichment was tested which is usually a high number compared to the number of terms for which even the uncorrected p-value reaches typical significance values. In both correction methods when determining the threshold of the null hypothesis rejection n appears in the denominator and forces correction with respect to the terms which were never realistically considered to be a 'discovery'.
There are other analysis frameworks for ontology term ranking and enrichment (Frost and McCray, 2012) which take structure of an ontology into account yet testing them is beyond the scope of this thesis.

3.3.5 *Outstanding issues*

Data-driven approach exercised in this chapter offers a new depth of the global level overview of diversification patterns in the full proteome and indicates significant differences in PSP with regard to origin of proteins and their key episodic selection periods. I was able to identify specific proteins of interest as well as discuss general trends in the context of the geological timeline and phenotypic correlates. However, the study did not offer support to the hypothesis about the link between shared selection events and common function. It remains unclear whether the hypothesis is untrue or methods employed here (such as ontology enrichment tests) are not

appropriate for the question. I postulate that this can be addressed by integrating other classes of data in the analysis (e.g. pathways and protein-protein interactions) and avoiding blanket enrichment analyses as they suffer from lack of power. In the following chapter I will implement these ideas further and demonstrate utility and unique insight brought by the temporal episodic selection modelling data.

4

EVOLUTION OF INTERACTING COMPLEXES OF PROTEINS

In the previous chapter I demonstrated one use of a large scale application of the methodological framework which was described in Chapter 2. There I focussed on data-driven unsupervised search for patterns and distinct temporal profiles of protein diversification. I identified key evolutionary differences between post-synaptic density and the full proteome, and studied the link between the common evolutionary timeline of a group of proteins and their shared characteristics through gene set enrichment analysis. Here, I use the same modelling results (temporal evidence for episodic selection pressure) but approach them from a different angle. I leverage protein-protein interaction data as well as pathway annotations to create systematic and structured protein groupings. Then I compare these structures, as well as individual proteins, based on their temporal selection pressure patterns, summary measures, and inter-protein distance metrics previously introduced in Chapter 3.

I address the methodological limitations of the broad enrichment analysis which became apparent in the previous chapter and instead of describing functional associations of temporal diversification clusters I use structure enforced by interaction data or pathway annotation to study how specific groups of proteins evolved differently to others, and in case of interaction data, how topological features of a protein in the interactome link to its evolution.

4.1 INTRODUCTION

Proteins do not act in isolation from one another, they bind with each other into complexes of various size and lifespan, they propagate signals, get modified by chemical reactions, and get recycled, also in a process involving multiple other proteins (Jones and Thornton, 1996). Dividing proteins into groups purely based on anatomical location following localised expression experiments serves only limited purpose as it does not take into account the wealth of interdependencies between proteins.
Focus on interactions between proteins and function of complexes of multiple pro-

teins highlights the functional relationships between proteins as opposed to anatomical co-occurrence. Furthermore, modelling protein interactions as edges of a graph allows for employing methodological approaches such as network analysis which adds informative structure to data.

4.1.1 *Interactions*

Protein interaction is a physical process occurring at a molecular level in which two (or more) proteins bind together using parts of their peptide chains for any amount of time. Multiple structural factors play a role in protein binding: amino acid residue preferences, hydrophobicity, electrostatic and shape complementarity, secondary structure, and size of accessibility area (see Jones and Thornton, 1996, for a review of biochemical principles governing direct protein binding). Although binding occurs between very specific regions of proteins (often described as binding domains), and often only in very specific circumstances, when modelling the behaviour of multiple proteins at one time it is common to abstract the complexity of this event to a binary fact of whether two proteins are capable of binding with each other. Thus, databases of protein interactions are simply lists of pairs of protein accessions of molecules which are capable of binding. Supplementary data such as experimental procedure or region annotation might be available and can be used for filtering purposes.

4.1.2 *Network analysis*

A large set of proteins can be analysed as a graph, where proteins are represented as nodes and interactions between them as undirected edges. This simplified interpretation of protein interactome allows me to employ a network analysis approach, where proteins are modelled as nodes and interactions as edges of a large undirected graph. It is an established computational framework for studying large sets of proteins and authors of large systematic studies in the field (Huttlin et al., 2017) claim high biological significance of network features of the interactome.

4.1.2.1 *Centrality*

In biological networks such as protein-protein interactions it is highly informative to identify proteins which are local hubs as their failure may break pathways. There are

multiple ways of measuring relative importance of a node in a graph, the simplest being its degree - number of edges attached to it, however, this measure is blind to any topological phenomena observable beyond a single level of direct connections. Thus measures such as betweenness centrality are used for this purpose. Betweenness centrality of a node V is defined as number of geodesics (shortest paths between any two nodes) traversing a node Freeman (1977).

$$c(V) = \sum_{A \neq V \neq B} \frac{\sigma_{AB}(V)}{\sigma_{AB}}$$

where σ_{AB} is the total number of shortest paths (geodesics) from node A to node B and $\sigma_{AB}(V)$ is the number of those paths that pass through V.
Essentially, the biological interpretation of this metric implies that the removal of a node with high betweenness centrality implies that a large number of pathways get disrupted (Jeong et al., 2001)

4.1.2.2 *Community detection*

In networks which model real-life phenomena, nodes often connect with each other preferentially, forming groups based on the principle of a node being more likely to connect with the nodes which are already connected with its neighbours. It is perhaps most intuitive if we imagine our friendships as edges of a graph with people as nodes - circles of friends can be interpreted as network communities (Ferrara, 2012). The objective of community detection procedures is to identify these groupings by only using the information in the graph topology.
Spin-glass (Eaton and Mansbach, 2012; Ispolatov et al., 2006) is one of many community detection algorithms (see Yang et al., 2016, for a review and benchmarks), and it has a record of being used in the context of protein interaction networks (e.g. Daraselia et al., 2007). The Spin-glass community detection algorithm is an iterative, non-deterministic approach borrowed from statistical physics. In this model, each vertex can be in one of N spin states, and the interactions between the vertices (i.e. the edges) specify which pairs of vertices would prefer to stay in the same spin state and which ones prefer to have different spin states. The model is initialised at random then simulated for a given number of steps; the resulting spin states define the community membership of the nodes (Eaton and Mansbach, 2012; Ispolatov et al., 2006).

The algorithm is implemented in many graph analysis libraries, including `igraph` `R` package.

4.1.3 *Pathways*

A complex of proteins responsible for a specific biological process can be described in a form of a pathway. Nodes, which represent proteins, or any other molecules, are connected with edges which can represent multiple processes such as temporary binding, self-association, secondary messenger signalling, phosphorylation, etc. This implies that a single pathway may contain the same protein in multiple nodes which stand for different phosphorylation states. Also, there tends to be a temporal direction enforced between members of the pathway thus creating linear paths of interactions between nodes. As mentioned in the previous chapter, in section 3.1.4, with appropriate annotation available, they can be modelled within an ontology framework, as more general pathways can be often broken down into several smaller pathways.

4.1.3.1 *Sources*

As discussed previously in section 3.1.3 the main sources of pathway annotation are KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2017) and Reactome (Joshi-Tope et al., 2005; Mi et al., 2017). Because of the nested, ontology structure of pathway annotation, Reactome is the preferred source of protein-to-pathway mapping.

4.1.3.2 *One-to-many mappings*

Many proteins may represent a single node because in a particular setting of this molecular function they all serve the same function in the pathway. This causes problems for any graph based analysis. Also, it should not be interpreted that if one of the proteins mapping to a given pathway node represents a desired property then the entire node has this property. Without any further information about the node we can only only assume all proteins mapped to it can perform its functional role in the pathway.

4.1.3.3 *Relationship diversity and molecular co-evolution*

Relationships between pathway members may represent various kind of effect one node has on the other, some being simple two protein bindings (Jones and Thornton,

1996), but others might include phosphorylation of another protein (Hunter, 1995), enzymatic activity (Newton, 1995), or building more complex multi-protein complexes (Houtman et al., 2005; Yang et al., 2008b). All classes of dependencies might enforce different constraints on protein co-evolution, however, without a systematic way of interpreting edges in an interaction database we are only able to extract generalised information. True protein co-evolution on a molecular level involves compensatory mutations at pairs of residues which aim at maintaining or improving specificity of a functional relationship between proteins (Pollock et al., 1999).
However, Talavera et al. (2015) question detection of true molecular co-evolution through covariation of the sequences, one of their hypotheses attributes observable correlation to independent substitutions amplified by the tree structure. Furthermore, Hakes et al. (2007) attempted to identify protein-protein interactions based on correlated evolution signatures yet this approach yielded worse results than using co-expression data, authors concluded that the observed correlated evolution could have been caused by an unmeasured latent variable exerting common selective constraint instead of true co-evolution. However, they also hypothesised about the co-evolution signal being available only at the interface regions of the proteins which could be too weak due to small relative size of these regions. Overall, molecular co-evolution in its strict meaning remains an elusive phenomenon to detect. Although on a conceptual level molecular co-evolution is naturally linked to a functional relationship between proteins any causality between observed measures of correlated evolution and protein relationships is harder to argue.

4.1.4 *Objectives*

This work is motivated by limitations of the previous chapter as well as existing work on pathways and protein-protein interactions. I propose integration of the functional data about protein behaviour introduced above with their evolutionary characteristics derived from my modelling work.

The first objective is to source interaction data and pathway data, then clean, quality check, and prepare them to be integrated with temporal diversification modelling data.
Then I aim to test the following hypotheses:

1. Interacting proteins evolved together, their diversification timelines are linked. This hypothesis can be further rephrased in a weaker non-causal and stronger causal way:
    - Proteins which interact are more likely to share selection pressure timeline (yet causality is unknown - perhaps unobserved latent variables explain it)
    - Changes in selection pressure in one protein can influence evolution of its interactors.
2. Position of a protein in the interactome network affects temporal pattern of its evolution.
    - Communities derived from network topology group together proteins with similar evolutionary profiles.
    - There is a link between the characteristics of protein's evolution and its relative importance in the interactome graph.
3. Pathways provide better grouping of proteins with respect to selection pressure events that interactome communities or anatomical divisions, which is an extension of work in the previous chapter addressing the link between common evolution and common function.
4. Pathways playing a role in synaptic function regulation diversified recently even if their members have a deeply conserved origin (as previously indicated in Emes et al., 2008; Emes and Grant, 2012)

Studying these hypotheses will build on the methods introduced in the previous chapter, such as summary measures of selection pressure timeline and groupings described in sections 3.2.1.1, 3.2.3, and 3.2.4. I will however also introduce new approaches more relevant to the other classes of data integrated through the course of the study.
Below, I present results for the interactome and for pathways in separate sections.

4.2 INTERACTOME RESULTS

4.2.1 *Data sources*

The database of interactions is an in-house resource (e.g. used by McLean & Sorokina, manuscript in preparation). First, newest versions of interaction records from Intact

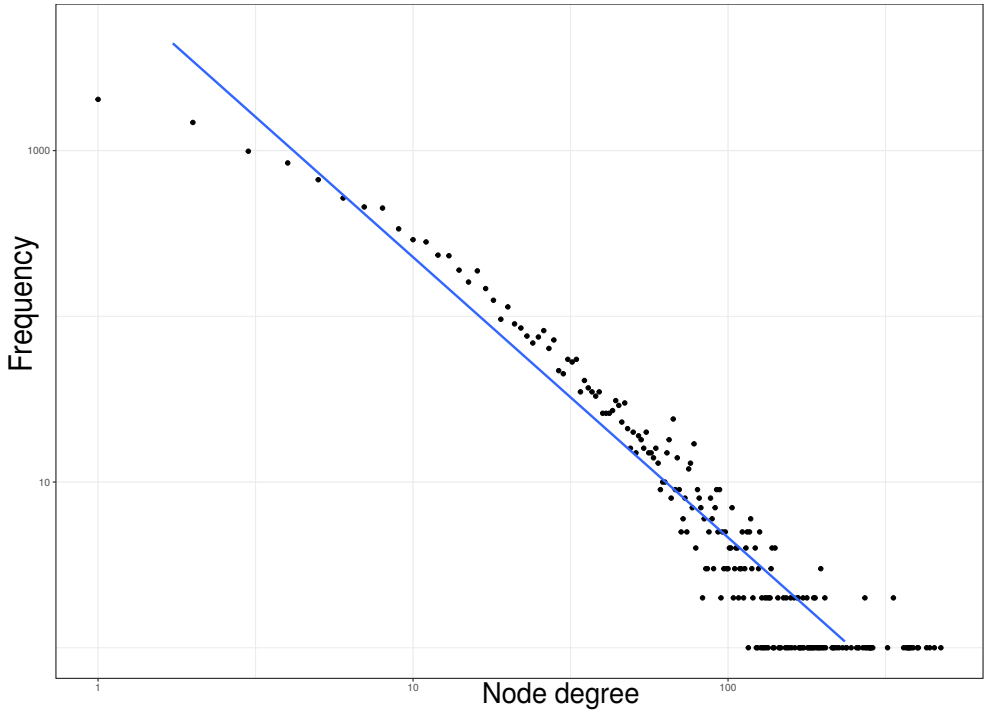


Figure 4.1: Relationship between node degree and frequency in the full human interactome, fit to an exponential decay function (or straight line on log-log scale) is indicative of scale-free property - typical for large biological networks.

(Kerrien et al., 2012; Orchard et al., 2014), Biogrid (Chatr-Aryamontri et al., 2017) and DIP (Xenarios, 2002) were downloaded. All identifiers were mapped to entrez ids, and data from all sources were integrated into one table. Each of the data sources listed above contains field informing about the type of evidence for the particular interaction, all in silico inferred interactions were excluded from the dataset (see von Mering et al., 2002, for a review of classes of interation evidence). Finally redundant duplicate interactions were eliminated.

An initial graph is generated only from proteins for which temporal selection pressure modelling data exist. After simplifying the graph by removing self-associations as well as multiple edges and subsequently extracting the largest connected component, the resulting interactome network consists of 11787 nodes and 71892 edges. Limiting the network to PSP genes and executing the same simplification procedure results in a post-synaptic interactome of 920 nodes and 2121 edges. It is a scale-free network, which is expected with a biological network of this size. It means that the frequency of nodes of a given degree decreases exponentially with the increase of

the degree (see Figure 4.1). A major deviation from this pattern and outliers could suggest issues with interaction annotation.

4.2.2 *Similarity of interactors*

The first hypothesis to be tested was investigating the relationship between temporal profile characteristics of pairs of interacting proteins. For a simple systematic test of a hypothesis that interacting nodes are more likely to be have similar diversification timeline I used used the same weighted cosine similarity of base matrix and smoothed matrix as in chapter 3 (see section 3.2.2.1 for details of calculations). Using the adjacency matrix of the graph as a guide for existence of a connection between proteins I compared the distances between all pairs of interacting proteins to distances between all pairs of non-interacting proteins yet there was observable difference in the systematic comparison over the entire proteome set (medians 1.232 vs. 1.258), nor was there one for the limited PSP, in fact median of similarities between connected nodes was slightly higher (1.329 vs. 1.244).

4.2.3 *Community effects*

In the following test I aimed to identify evolutionary effects in the interactome communities. However, I performed this analysis on the post-synaptic interactome to achieve more biologically relevant clusters. 24 communities were identified using the spin-glass procedure (see section 4.1.2.2, 13 of them had at least 20 nodes and only these communities were considered further (limitation was imposed to retain sufficient sample for summary analyses - summary measures for small sets can be misleading). I analysed communities' internal cohesion and differences between them using the intra-community distance metric (section 3.2.2.1 as well other measures derived from the temporal selection profile vector did not reveal significant and systematic effect of communities evolving synchronously (see Table 4.1). To put the numbers in the table into perspective, in the entire PSP subgraph of full human interactome, median distance between pairs of proteins is 1.244; median relative diversification window is 0.478, its IQR is 0.336; median relative total positives is 0.136, and its IQR is 0.136.

Communities number 1, and 2 are characterised by the lowest median distance between community member, community number 1 has a particularly high proportion of early origin proteins (predating nervous system development) and low proportion of recently positively selected genes. On the other hand community number 2 displays opposite pattern of these two features. Members of these two communities are listed in Table 4.2. Despite these differences, when measured on corrected timeline measures, the two communities do not differ much. In fact, most communities do not deviate much on these measures; however, some, such as numbers 9 and 17, show substantially less variance in diversification window; similarly 10 and 12 are characterised by low variance of total positives. In both cases this did not seem to have an effect on the similarity of their timelines as measured by median inter-member distance.

Overall, on a systematic level communities did not group together proteins of similar diversification profile, with the notable exception of two communities (1 and 2 in Table 4.1); members of these two communities present interesting targets for further validation.

Table 4.1: Communities with at least 20 members in PSD interactome and their evolutionary profile characteristics.

		Distance		Diversification window		Total positives		Pre-NS proteins		MRP=1 proteins	
Community	Members	median	max	median	IQR	median	IQR	number	fraction	number	fraction
1	56	1.097	2	0.478	0.348	0.136	0.127	34	0.607	5	0.089
2	24	1.058	1.965	0.5	0.398	0.146	0.103	5	0.208	3	0.125
5	71	1.367	1.994	0.435	0.327	0.13	0.104	53	0.746	7	0.099
6	41	1.247	2	0.435	0.334	0.13	0.13	27	0.659	3	0.073
8	31	1.264	1.999	0.409	0.266	0.13	0.111	15	0.484	2	0.065
9	53	1.217	2	0.409	0.204	0.13	0.13	21	0.396	3	0.057
10	64	1.271	2	0.409	0.312	0.13	0.089	42	0.656	5	0.078
12	22	1.423	1.999	0.435	0.334	0.113	0.078	14	0.636	0	0
13	30	1.188	2	0.435	0.351	0.174	0.14	15	0.5	1	0.033
17	49	1.252	2	0.381	0.206	0.13	0.104	27	0.551	2	0.041
19	69	1.316	2	0.435	0.323	0.143	0.126	28	0.406	5	0.072
20	82	1.133	2	0.435	0.334	0.133	0.095	43	0.524	2	0.024
22	98	1.351	2	0.435	0.364	0.136	0.14	45	0.459	8	0.082

Table 4.2: Members of selected graph communities; * protein with evidence for most recent episodic positive selection in h.sapiens branch, †protein with orthologs in organisms predating nervous system development.

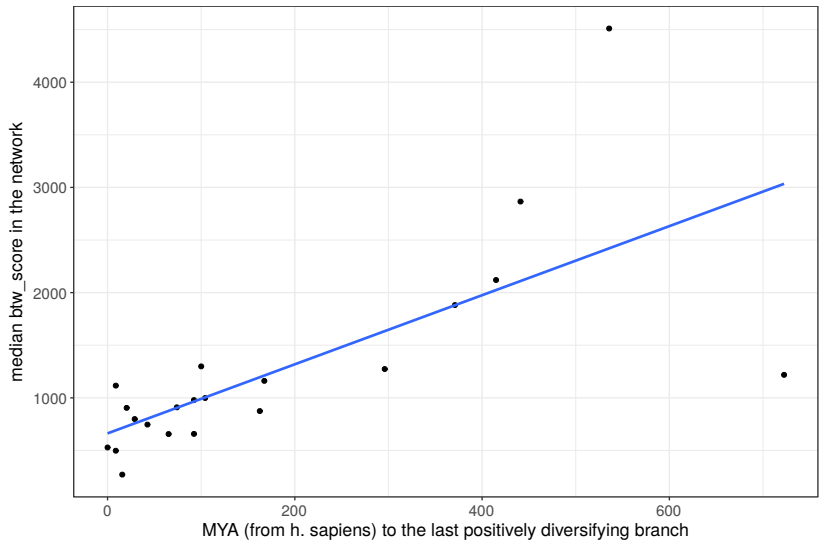
Community	List of members
1	SYNJ1*†, DNM2*†, SORBS1*, TACC1*, PALM*, ITSN1†, AP2A1†, HIP1†, SEC24C†, RPL4†, ATP6V1G2†, TIMM50†, HSPD1†, IPO5†, CORO1C†, PYGM†, STRAP†, GLUL†, DNM1†, ARPC4†, KIF5A†, AP2B1†, WASL†, ACTR2†, ACTR3†, TARSL2†, ARPC2†, CORO1B†, NME1†, MYO1E†, ATP6V1E2†, DPYSL2†, AP2A2†, SUCLA2†, DPYSL3†, DPYSL4†, PYGB†, AMPH, SH3GL1, HTT, SH3GL3, PEX11B, SNX9, CLASP1, BIN1, WDR91, DLGAP4, PACSIN1, GSTM3, GBAS, PACSIN2, ANKRD24, WIPF2, ADD3, RAPH1, PPP1R21
2	RASAL2*†, PLEKHA5*, SPTB*, DNAJA3†, DCLK1†, AGAP1†, ENO3†, SPTAN1, DLG4, SPTBN1, KTN1, NOMO1, STAT1, SHANK1, GRIN2A, GRIN2D, SH3PXD2A, LRP1, NDUFA9, EVL, SIPA1L1, LIMA1, PLEKHA6, HADH

4.2.4 *Centrality relationship*

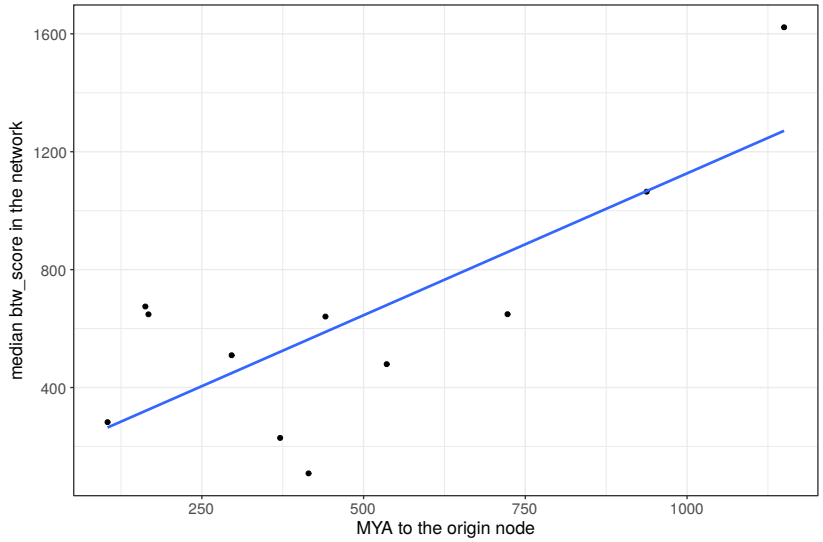
Having failed to capture clear effects based on groupings of interactors I turned to topological properties of individual properties, specifically their relative importance for the graph, the node centrality (see section 4.1.2.1).
Although most centrality measures correlate with each other to a certain degree, betweenness centrality is easy to interpret as a biological phenomenon; therefore it was selected as the measure of importance for proteins in the interactome.

I observed correlation between distance to the most recent significantly positively selected branch (denoted as MRP, see section 3.2.3 for an explanation of the measurement). The effect becomes clearer when datapoints are binned by levels of MRP (see plot in Figure 4.2a. However, I also observed a similar correlation with the origin of the protein (methodology of measuring protein origin explained in section 3.2.1.1) which is shown in Figure 4.2b; therefore in light of dependency between these two measures (section 3.2.3.1) I tested the relationship between node centrality and the relative length of diversification window (see section 3.2.4 for details of how the mea-

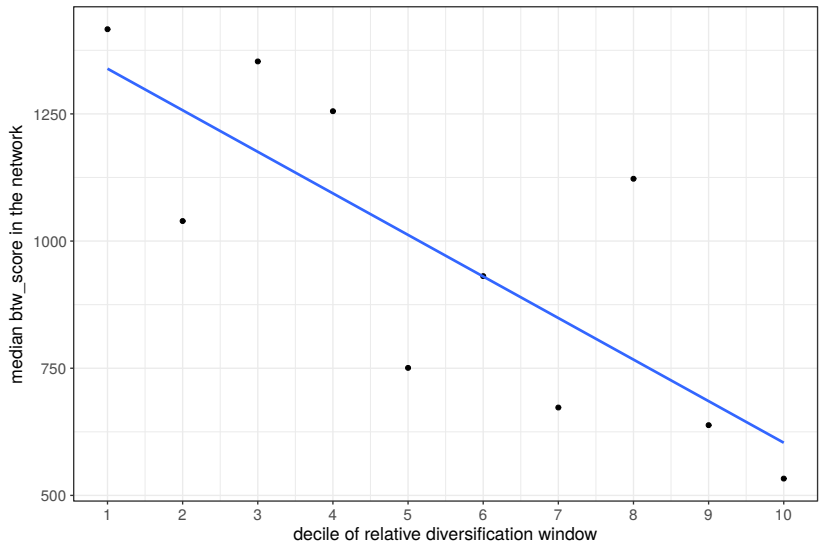
surement is derived). I found a negative trend in this comparison, i.e. longer relative diversification window translates to lower centrality of a protein (see Figure 4.2c). These trends may suggest that the evolutionary history of an individual protein may be related to its topological position in the molecular interaction graph rather than its membership in groupings such as ontology terms or communities.



(a) Proteins binned by their MRP.



(b) Proteins binned by their origin point.



(c) Proteins binned by decile of the relative length of a diversification window.

Figure 4.2: Relationship between protein centrality and temporal evolution measures: MRP (see section 3.2.3), origin point (see section 3.2.1.1), and relative length of a diversification window (see section 3.2.4). Median betweenness centrality per bin on y-axis in all three plots.

4.2.5 *Hub effect*

The trend observed across the plots in Figure 4.2 motivated an attempt to test mechanistic explanation of a relationship between interactions and positive episodic selection.
I hypothesised that there is a trend of a radiating episodic selection impulse which originates at a hub and is spread through edges of the network to nodes of lower centrality, which could be interpreted biologically as adjustment of proteins in response to a change in a protein which they interact with but is more important in the network.

4.2.5.1 *Hypothesis*

Here I test two hypotheses:

1. An episodic positive selection event in a central hub selection leads to episodic positive selection in its interactors which are lower in centrality. Although I hypothesise positive selection spreads with a delay (i.e. it would appear in the first degree interactor node after one timestep, then in the second degree interactor after 2 timesteps) it is to be determined whether it is possible to capture that precise aspect of the effect with the low temporal resolution of nodes in the tree.

2. The appearance of a central protein induces episodic diversifying pressure in its interactors of lower centrality. This is similar to the first hypothesis but the event triggering the spread of positive selection is the appearance of a highly central node among other, already existing, nodes of lower centrality.

4.2.5.2 *Testing methodology*

For top hubs of the full interactome I identified chains of two consecutive interactors in such a way that:

- At each level of a chain $K \in \{0, 1, 2\}$ (where $K = 0$ is the hub itself) all proteins have lower centrality score than the protein to which they are connected at level $K - 1$

- For each hub none of its $K = 1, 2$ interactors are also connected to the other hubs

- For each hub none of its $K = 2$ interactors are also connected to another $K = 1$ interactor node of this hub

Figure 4.4 illustrates how chains are selected. Then, when testing the first hypothesis, for each hub I aggregate results across all chains for each branch with significant episodic positive selection at time T. I count the frequency of significant positive selection on branches $T, T+1, T+2$ for interactors $K = 1, 2$, and compare these frequencies to baseline frequencies on these particular branches. Annotation of branches is explained in Figure 4.3. Baseline frequencies are simply frequencies of significant positive diversification on all branches of root human path (not to be confused with the MRP frequencies).
When testing the second hypothesis, the procedure is similar, but T is the origin point of the protein. Also, chains are limited to interactors $K = 1, 2$ which originated at least 2 branches prior to the hub, and hubs are limited to proteins with origin point of 21 or more recent. In both cases the top 100 qualifying hubs are considered.

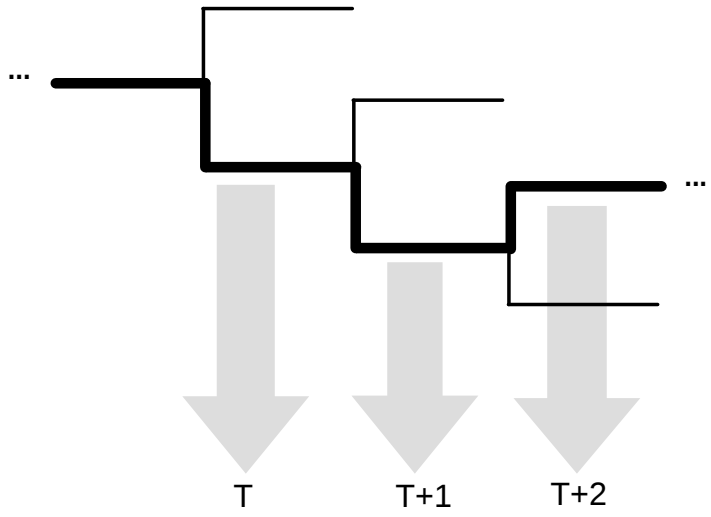


Figure 4.3: Branches annotation in testing hub-chain effect shown on a fragment of the phylogenetic tree. On the highlighted path from root to human branch T is the oldest one, $T+1$, and $T+2$ are the subsequent more recent branches.

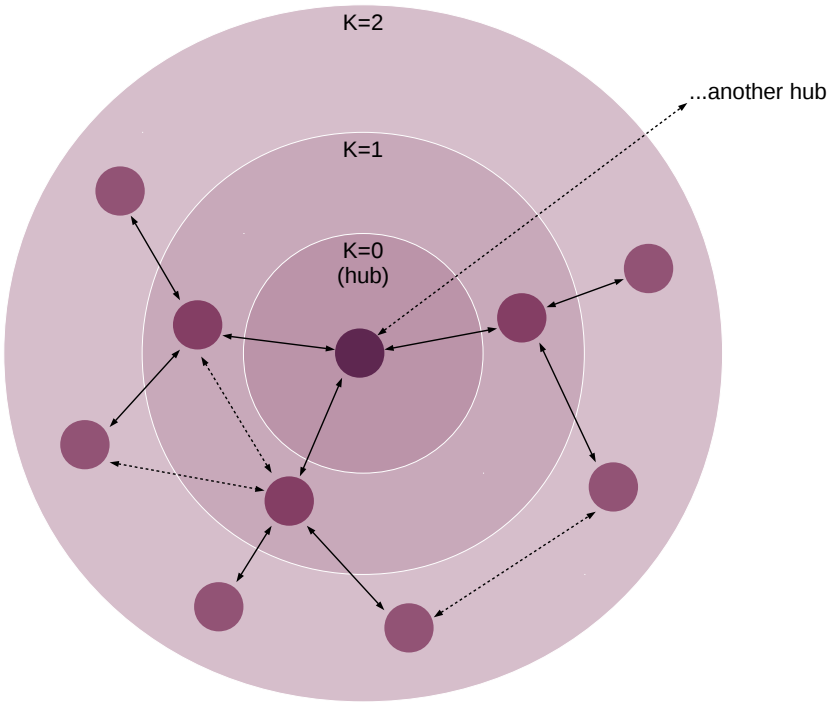


Figure 4.4: Methodology of testing the hub-chain effect, solid lines are the necessary connections between elements, dashed lines are optional but allowed connections as long as the principle of decreasing centrality is maintained (see section 4.2.5.2) , only nodes within $K = 2$ circle form the group of chains for a given hub, other hub listed in top-right corner would be a source of another group of chains extending from it.

4.2.5.3 *Results*

I observed no effect for all positive branches of a hub aggregated together (Table 4.3). However, there was an effect for the oldest positive branch in a hub (Table 4.4), as well as a very clear effect of the appearance of a hub node in the pre-existing neighbourhood (Table 4.5).
Temporal aspect of the effects remains unclear, based on the magnitude of difference against baseline frequencies, and the level of statistical significance of the test, branch $T + 1$ seems to be the most affected; for the appearance effect in the second interactor ($K = 2$) it is the branch $T + 2$ which may suggest a secondary effect of positive diversification in $K = 1$ interactor at branch $T + 1$.

Overall, these results lend support to the novel hypothesis of the radiating effect of positive selection events which could explain effects described earlier in section 4.2.4 yet further investigation into the specific aspects of the proposed mechanism is necessary.

Table 4.3: Centrality effect results for all positive branches, difference between frequency of significant positive episodic selection in a set of interactors and baseline frequencies aggregated over all hubs and all positive branches. See section 4.2.5.2 for details of how these numbers were generated. In parentheses p-values of t-test against true $mean = 0$.

Interactor	Branch T	Branch T + 1	Branch T + 2
First (K = 1)	$0.0025_{(0.22)}$	$0.0032_{(0.16)}$	$-0.0002_{(0.52)}$
Second (K = 2)	$0.0020_{(0.04)}$	$0.0015_{(0.07)}$	$-0.0005_{(0.71)}$

Table 4.4: Centrality effect results for oldest branches, aggregated over the oldest positive branches for each hub. See section 4.2.5.2 for details of how these numbers were generated. In parentheses p-values of t-test against true $mean = 0$.

Interactor	Branch T	Branch T + 1	Branch T + 2
First (K = 1)	$0.0080_{(0.08)}$	$0.0161_{(0.01)}$	$0.0095_{(0.06)}$
Second (K = 2)	$0.0067_{(0.002)}$	$0.0072_{(0.001)}$	$0.0050_{(0.005)}$

Table 4.5: Centrality effect results for appearance points , aggregated over all hubs and their origin points. See section 4.2.5.2 for details of how these numbers were generated. In parentheses p-values of t-test against true $mean = 0$.

Interactor	Branch T	Branch T + 1	Branch T + 2
First (K = 1)	$0.0394_{(<0.001)}$	$0.0426_{(<0.001)}$	$0.0336_{(0.006)}$
Second (K = 2)	$0.0155_{(<0.001)}$	$0.0337_{(<0.001)}$	$0.0389_{(<0.001)}$

4.3 PATHWAYS RESULTS

Following Emes et al. (2008) I adopted the approach of tracking protein complexes through their evolutionary history, however, with two modifications. First, I define groups through their functional co-dependencies within curated pathways. Second, I analyse timeline of diversifying selection pressure on top of the point of origin of a protein (protein origin is also part of the analysis).

4.3.1 *Pathway selection*

I selected pathways from Reactome (Joshi-Tope et al., 2005) and acquired graph structures of the pathways from the Reactome plugin in `Cytoscape` (Wu et al., 2014; Shannon et al., 2003). Reactome is a hierarchical ontology, but the depth at which I picked pathways was arbitrary, however, for most of them I also added pathways represented by descendant nodes of the broader pathway (sub-pathways). However, I did not do so exhaustively, the motivation was to investigate whether sub-pathways can represent different episodic diversification profiles from its parent pathways and from its sibling pathways. Chosen pathways describe a selection of molecular processes implicated in complex neural function introduced in Chapter 1:

- **Neurotransmitter signalling** - see section 1.2.1, with a sub-pathway:
  - **Neurotransmitter release cycle**
- **GPCR signalling** - see section 1.2.3.2
- **Translation** - see section 1.2.3.3, also, a descendant node of this pathway:
  - **Translation Initiation**
- **Vesicle mediated transport** - see sections 1.2.1 and, also, two sub-pathways from the ontology tree:
  - **Vesicle Binding Uptake**
  - **Vesicle Membrane Traffic**
- **Axon Guidance** - see section 1.2.2, and two of its many sub-pathways, each focussed on a single extracellular signalling molecule:
  - **Axon Guidance RET**
  - **Axon Guidance EPH**

- **NGF Signalling** - see section 1.2.2
- **Innate Immune System** (Janeway and Medzhitov, 2002) - it is an unrelated, background pathway to test the possibility that there is an inherent bias of selection pressure timeline representation for proteins assigned to Reactome pathways in general.

Member counts and interaction counts as well as annotations which will be used throughout tables and figures in this chapter are summarised in Table 4.6.

Table 4.6: Tested pathways, numbers refer only to proteins for which there is temporal selection pressure data with at least one branch of significant positive episodic selection, interactions are sourced through Reactome plugin in `Cytoscape`

Pathway Name	Abbreviated name	Proteins	Interactions
Neurotransmitter signalling	neurotransmitter	109	668
Neurotransmitter release cycle	releasecycle	42	279
GPCR signalling	signalingGPCR	905	15219
Translation	translation	153	7345
Translation Initiation	translationinit	82	4645
Vesicle mediated transport	vesicle	533	11494
Vesicle Binding Uptake	vesicleBindingUptake	38	39
Vesicle Membrane Traffic	vesicleMembraneTraffic	497	11455
Axon Guidance	axonguidance	455	4088
Axon Guidance RET	axonguidanceRET	203	2227
Axon Guidance EPH	axonguidanceEPH	77	953
NGF Signalling	NGFsignalling	362	2527
Innate Immune System	innateimmune	999	1314

4.3.1.1 *Differences in evolutionary profiles between pathways*

First, I started with the simplest evolutionary measure and studied origin of proteins in pathways. Figure 4.5 presents a visualisation of the effect of variable representation of different bands of protein origin. For these heatmaps protein origin (the deepest node with existing ortholog) is divided into 4 bands as described in section 3.2.1.1). Each pathway (as well as the full proteome for reference) is presented in

one column, and for each of them proteins in each of the bands are counted to create Figure 4.5a. Then counts are normalised per column to create Figure 4.5b. Finally, I am interested in how each protein deviates from the background distribution of full proteome, hence in Figure 4.5c I divide the frequencies in each pathway's column by reference frequencies for the full proteome (rightmost column in Figure 4.5b). I also test whether the difference in counts is significant with Chi-square test corrected for multiple comparisons (FDR).

Protein origin representation offers a way of differentiating any groups of proteins, in this case pathways. However, as I demonstrated in Chapter 3 protein origin is not sufficient to explore evolution of groups of proteins in a meaningful way, hence I turned to the most recent diversification measure (MRP) (see section 3.2.3 in the previous chapter for explanation of the measure). A visualisation of protein frequencies in different MRP categories as well as ratios of these frequencies compared to background data was produced in the same way as described above for Figure 4.5; the results are shown in Figure 4.6.
Most neural pathways studied here are characterised by a moderate over-representation of the early MRP category and an under-representation of the recent MRP category compared to the background. GPCR signalling and innate immune system pathways are comparable to the background across all three categories. Translation and its sub-pathway of translation initiation are strongly overrepresented in the early MRP but strongly under-represented in the more recent two categories, finally, vesicle binding uptake (a sub-pathway of the vesicle pathway), is under-represented in the earliest category and moderately overrepresented in the intermediate category (yet none of the tests for this pathway are significant).

However, taking into account representation of levels of protein origin as showed in Figure 4.5, I performed another analysis of MRP representation but only for proteins from the deepest origin category, predating nervous system (pre-NS) (see Figure 4.7). This view of most recent positive diversification pattern extends hypotheses proposed in recent studies about post-synaptic proteins being present in early eukaryotes by demonstrating that in pathways such as GPCR signalling, many early origin proteins have not completed their diversification under positive selection pressure until very recently (primate branches). It also demonstrates that recent diversification observed

in Figure 4.6 is not exclusively an effect of proteins of recent origin.

Also, the innate immune system pathway I had picked for comparison evolved in sync with the full proteome background, both with respect to the distribution of origin of its members and most recent positive diversification branch, which supports validity of this grouping and demonstrates how variability between profiles in reference to the full proteome (such as in Figures 4.5c, or 4.6b) is a meaningful measure.

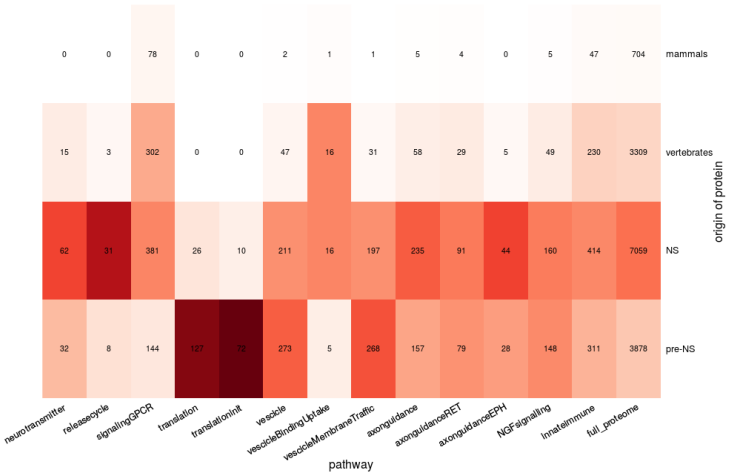
4.3.2 *Deep conservation & recent diversification*

Proteins with orthology which can be tracked back to primitive organisms (deep conservation) but with evidence for recent diversification are particularly interesting in the context of the synapse, as general function of a synapse emerged early, yet complex cognition developed relatively recently. Clearly there are a lot of synaptic proteins with ortholog predating nervous system as such (see Figure 3.2), but only some of them experience positive selection in the most recent branches of the root-human path. In the context of anatomically delineated PSP protein grouping I already presented sets of proteins which are potentially interesting because of their extreme temporal selection pressure profiles in the Table 3.4.

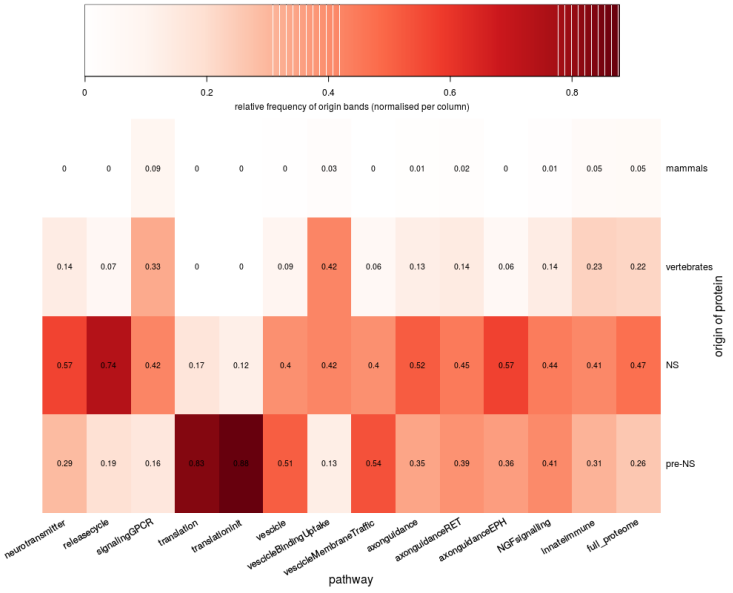
Here, in the context of pathways related to synaptic function Pre-NS column in Table 4.8 lists proteins which have known orthologs in organisms without nervous system yet with evidence for diversification in the *H. sapiens* branch of the phylogenetic tree (see Figure 3.3). These proteins are partly responsible for the top row of the heatmap in Figure 4.7 although relative frequencies of recently diversifying proteins are lower than in the full proteome.
Also, I identified potentially interesting targets for further research based on extreme values of timeline measures similar to Table 3.4 but divided by pathway - Table 4.9.

Using normalised measures (introduced in section 3.2.4) I observed a qualitative difference in distributions of the relative diversification window for two of the pathways, with probably the most extremely opposite patterns of both origin and MRP (in Figures 4.5 and 4.6 respectively) - GPCR signalling and translation. In Figure 4.9 I contrasted these pathways. Although the dominant peak of both distributions falls on the same value, in the GPCR pathway there is a noticeable shift of weight to-

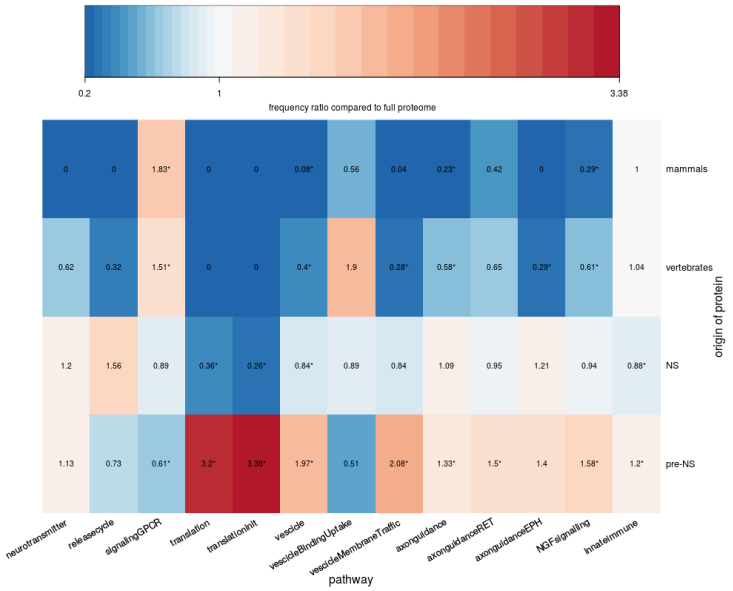
wards higher values. Also, higher variance can be observed for the GPCR pathway - its members exhibit more varied patterns of diversification within the set due to the pathway multi-functionality as opposed to the translation pathway with ubiquitous yet much more narrowly defined function.



(a) Counts of proteins

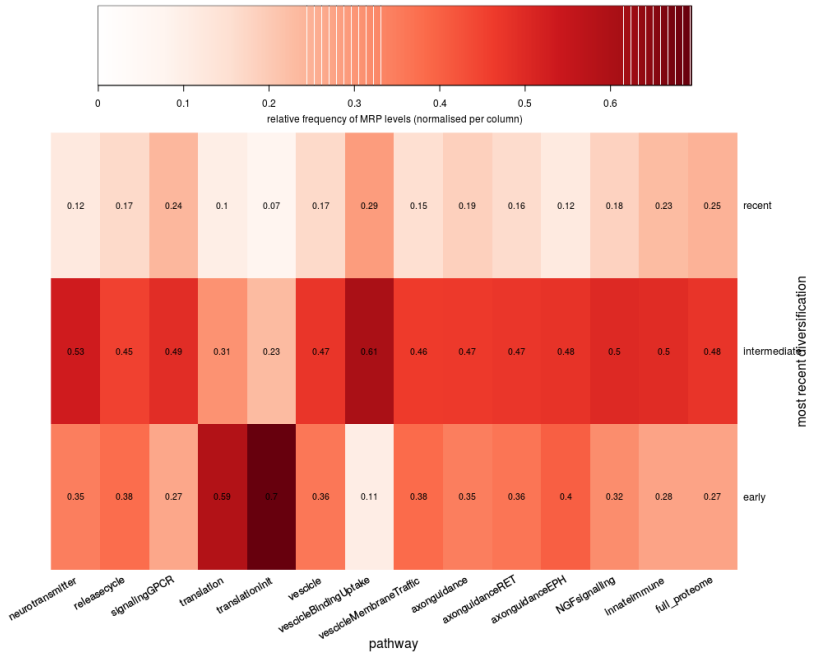


(b) Counts from the plot above normalised by column to frequencies

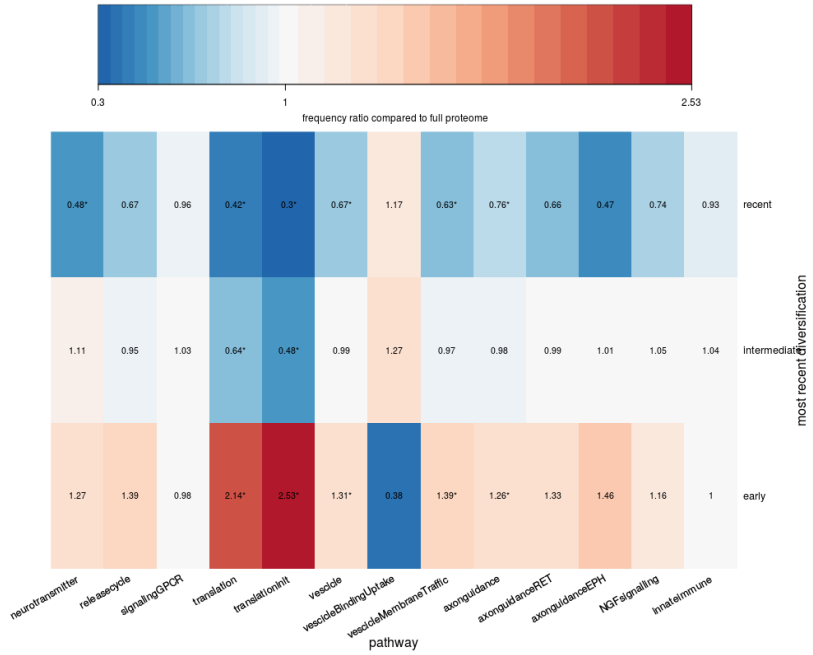


(c) Relative freqeuncies for each pathway divided by background frequencies for full proteome

Figure 4.5: Origin of proteins in selected pathways compared to reference representation of origin periods in the full human proteome. Process of deriving values in heatmaps described in section 4.3.1.1. Note very early origin in translation pathways, predominantly early in most synaptic pathways, and late origin in GPCR signalling.

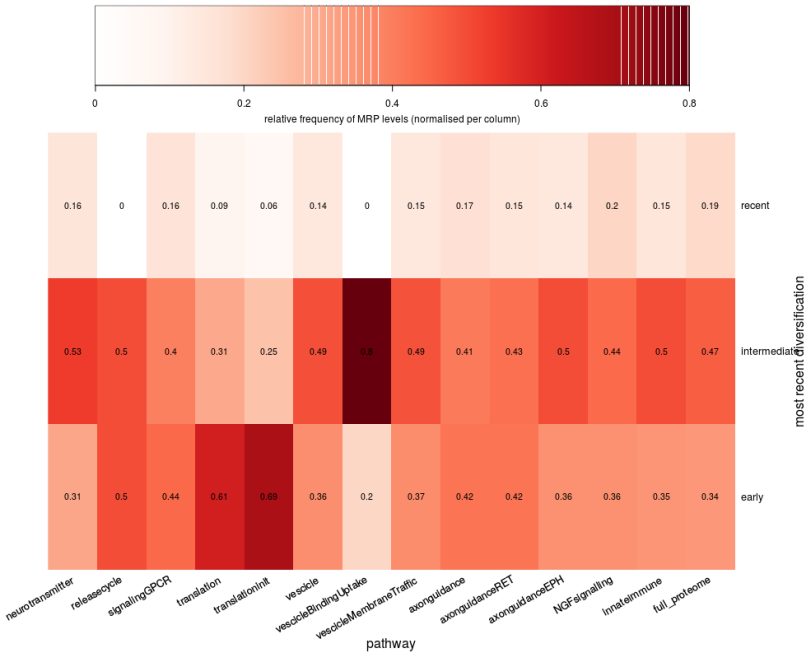


(a) Counts in each cell normalised by column to frequencies

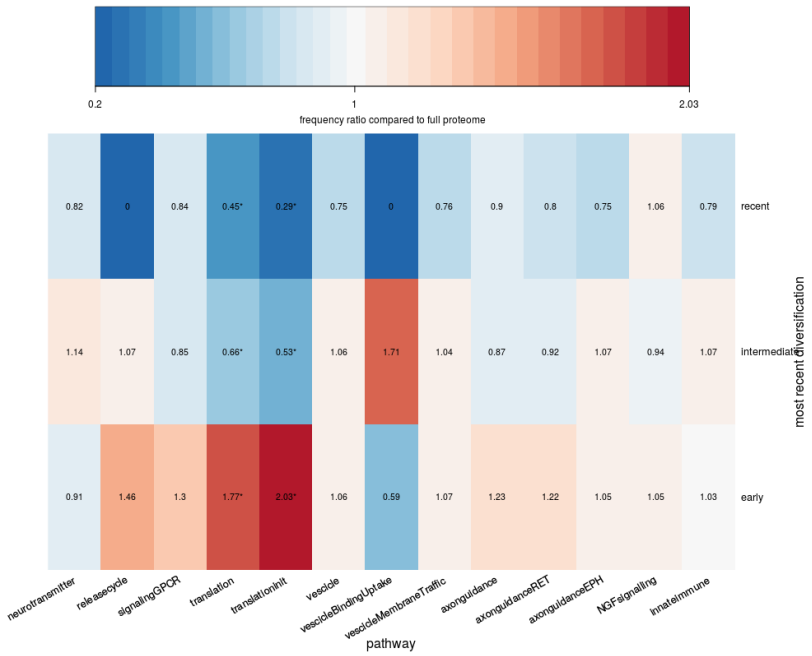


(b) Relative frequencies for each pathway divided by background frequencies for full proteome

Figure 4.6: MRP in selected pathways compared to reference representation of MRP in the full human proteome. Process of deriving values in heatmaps follows the same procedure as in Figure 4.5, heatmap with cell counts omitted. Limited evidence for recent MRP in most synaptic pathways apart from GPCR signalling.

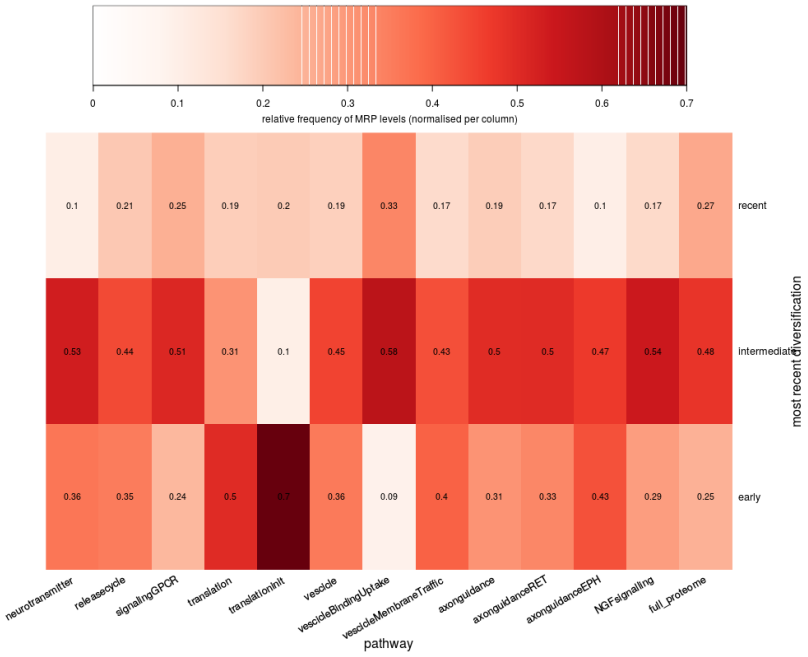


(a) Counts in each cell normalised by column to frequencies

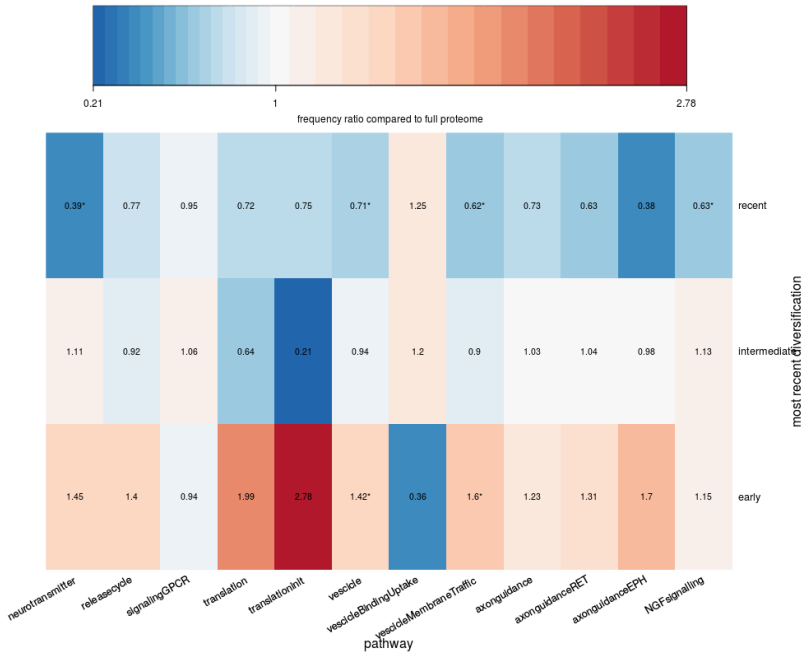


(b) Relative freqeuncies for each pathway divided by background frequencies for full proteome

Figure 4.7: MRP in selected pathways compared to reference representation of MRP in the full human proteome restricted to proteins with origin in eukaryotes preceding nervous system development (pre-NS group from section 3.2.1.1). Process of deriving values in heatmaps follows the same procedure as in Figure 4.5, heatmap with cell counts omitted. Compare to Figure 4.6 - less recent MRP but the overall pattern prevails.

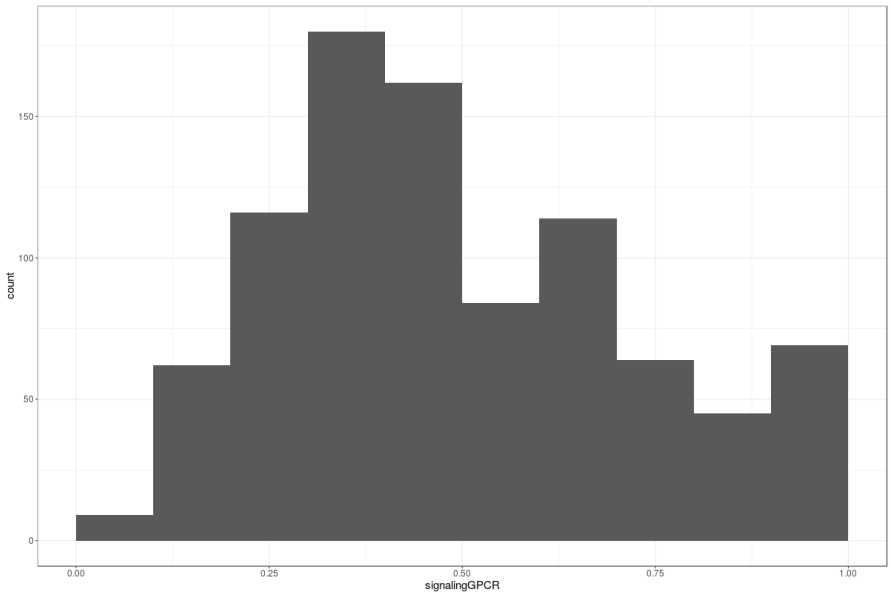


(a) Counts in each cell normalised by column to frequencies

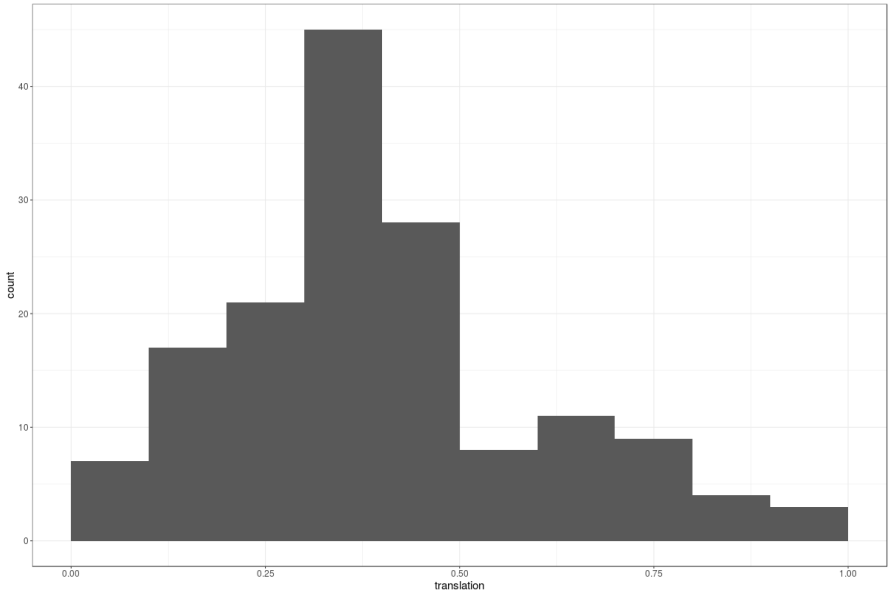


(b) Relative freqeuncies for each pathway divided by background frequencies for full proteome

Figure 4.8: MRP in selected pathways compared to reference representation of MRP in the full human proteome restricted to proteins with origin in eukaryotes which have nervous system development (Mammals, Vertebrates and NS groups from section 3.2.1.1) . Process of deriving values in heatmaps follows the same procedure as in Figure 4.5, heatmap with cell counts omitted. Compare to Figure 4.6 - general patterns remain even for translation pathway which is predominantly composed of pre-NS proteins (Figure 4.5a).



(a) Relative diversification window of GPCR signalling members



(b) Relative diversification window of Translation members

Figure 4.9: Distribution of timeline measures for two pathways: GPCR signalling and Translation.

4.3.3 *Pathways as graphs*

Here I aimed to reconcile some of the network analysis observations from section 4.2 with a different way of thinking about relationships between proteins within pathways. I investigated a similar effect to the hub effect for protein-protein interaction data described in section 4.2.5.

The main issue when analysing protein pathways as graphs is that one functional node (i.e. one node in a pathway diagram) can map to multiple different proteins (usually paralogs of the same family) and in fact they can display different diversification timelines (see Figure 4.10) ; normally the only evolutionary feature they share is their origin point (as if they are paralogs from one family they share the earliest ortholog). For the purpose of the goals of my study this limited my analysis to testing positive selection response to appearance of a pathway node.

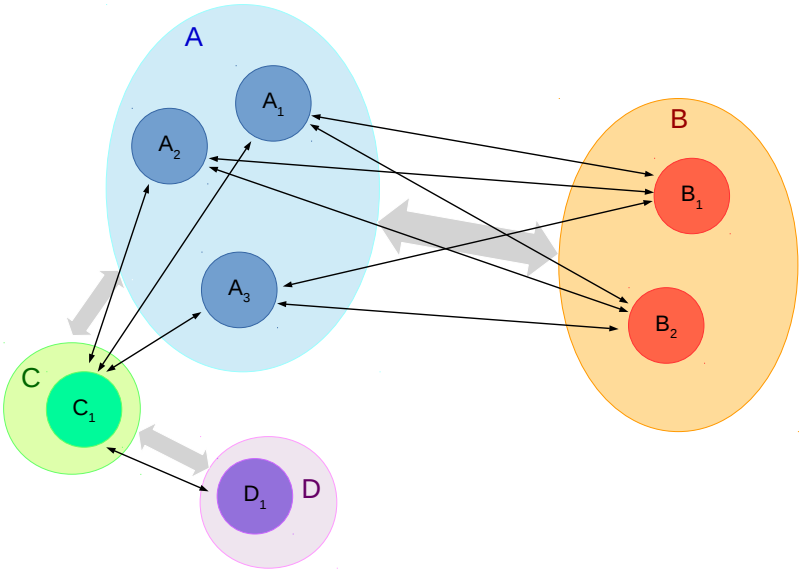


Figure 4.10: Groups of equivalent nodes in pathway graphs. Each aggregated node can resolve to a single protein (such as C, D) or multiple proteins (A, and B), normally paralogs within the same family. In the aggregated graph all 6 edges between A_1, A_2, A_3, B_1, and B_2 translate to a single edge between nodes A and B which is more informative of the actual dependency between elements of the pathway.

Pathway links sourced from `Cytoscape` Reactome plugin (Wu et al., 2014; Shannon et al., 2003) are the expanded links between individual proteins so, in the first step, I transformed these graphs to aggregated graphs (see Figure 4.10) while filtering only proteins for which temporal selection data is available (see section 3.2.1).
Then betweenness centrality measures were obtained for nodes of aggregated graphs. In order to obtain a sufficient sample for systematic inference I could only test pathways with enough new, and relatively central proteins appearing more recently than origin point of 22 while their interactors are already present (i.e. appeared at origin point of 22 or earlier) (see section 3.2.1.1 for explanation of protein origin metric).

Over all protein pathways early origin was highly overrepresented among most central nodes to a much higher degree than in the protein-protein interactome. Only the GPCR signalling pathway satisfied the criteria above (see Figure 4.5 for a comparison of origin points for different pathways), though even in this pathway the top 5 most central aggregated nodes had origin of 23 (predating nervous system).
For the GPCR pathway the step of graph aggregation transformed a protein graph of 1096 nodes and 15024 edges to an aggregated graph of 205 nodes and 718 edges.

Further limitation of this analysis was taking only a single level of interactors into account, as building two level chains within a network so tightly coupled is difficult as hubs are often very close to each other and encroach on one another's circle of influence (see Figure 4.4).
Each first level neighbour of the hub node was expanded to all its protein nodes, and only hubs with more than 5 expanded interactor nodes (i.e. protein nodes) were retained. From that point the analysis follows the same procedure as in the interactome, i.e. for each hub I aggregate results across all chains. I count frequency of significant positive selection on branches $T, T+1, T+2$ for interactors $K = 1$ (where origin point of the hub is T), and compare these frequencies to baseline frequencies on these particular branches (see section 4.2.5.2 for comparison).
The baseline frequencies were frequencies of significant positive diversification on all branches of root - human path for all members of the pathway for which data are available. This made detection of pathway-wide effects of introduction of a new hub impossible but it was the most direct way of testing local neighbourhood effects, in order to test pathway-wide effects a different baseline would be needed. The results

are summarised in Table 4.7.

Similar to the results for the interactome in Table 4.5 the temporal aspect of the effect remains unclear; here, an over-representation of positive episodic selection was only observed for branches T, T + 1 but not T + 2 (where T is the origin point of the hub node).
These results extend and generalise the hub-chain hypothesis tested previously in section 4.2.5. Not only physical interaction between proteins can propagate positive selection in response to a new hub protein, pathway links between the proteins can fulfil this role too.

Table 4.7: Centrality effect results for appearance points in GPCR signalling pathway, aggegated over all qualifying hubs and their origin points. See sections 4.2.5.2 and 4.3.3 for details how these numbers were generated. In parentheses p-values of t-test against true mean= 0.

Interactor	Branch T	Branch T + 1	Branch T + 2
First (K = 1)	$0.0868_{(<0.001)}$	$0.0739_{(0.003)}$	$-0.0003_{(0.5)}$

4.4 DISCUSSION AND CONCLUSIONS

This study was set out to explore temporal diversification patterns in relation to protein's topological features in the interactome as well as their role in functional pathways (as annotated in Reactome ontology).
I sourced interaction data, used common network analysis tools to group and order proteins according to their topology. I found the relationship between protein's centrality and measures such as protein origin, most recent positive diversification and length of diversification window. I tested whether the interaction between two proteins of varying centrality may be related to dependency between their episodic selection events as well as reaction to protein emergence among the pre-existing proteins.
Further to that I uncovered how selected important synaptic pathways represent a wide variety of temporal profiles of selection profiles with a big emphasis on early emergence but very recent diversification of their members which distinguishes them from characteristics of the full human proteome as well as anatomically defined PSP.

Overall, the approach of grouping proteins by their molecular functional role, then exploring their evolutionary history, adopted in this chapter proved more productive in explaining global, as well as PSP-specific, evolution patterns compared to the approach of exploring functional correlates of groupings driven by modelling data alone in Chapter 3.

I will discuss specific biological findings related to pathways and their members first, before more systematic effects based on interaction data as well as analysis of graph structure in pathways.

4.4.1 *Pathways*

In Figures 4.5 and 4.6 it can be observed that pathways related to synaptic function considered in this study fall into three types of diversification profiles: (1) highly conserved - early origin, mainly early MRP, (2) intermediate - mainly early origin, mainly early and intermediate MRP, and (3) recent diversification - origin spread in time, and substantial intermediate and recent MRP. In most cases sub-pathways follow approximately similar profile to their parent pathways. Vesicle binding uptake can be considered an interesting exception here, as it appears to have more recent origin proteins and higher rate of intermediate and recent MRP compared to its parent pathway, its recency of origin and MRP is even pronounced than in the GPCR pathway, it is hard to generalise from this finding as this pathway has very few members (which is illustrated by non-significant differences in frequencies compared to full proteome background).

Unsurprisingly the translation pathway and its sub-pathway of translation initiation were found to be the most conserved. Even correcting for dependency between protein origin and MRP by analysing exclusively proteins of pre-nervous system origin in Figure 4.7 there were significantly fewer proteins in the translation pathway which diversified until recently (recent and intermediate categories) than in the entire proteome, as well as compared to other pathways studied here.

The majority of pathways related to synaptic function followed the intermediate profile characterised by dominant origin of proteins in NS and pre-NS categories, in which they generally matched or exceeded frequency in the full proteome. There were

considerably fewer proteins in these pathways with vertebrate or mammalian origin compared to the full proteome. Considering the most recent evidence for episodic positive pressure these pathways have generally more early MRP proteins, similar fraction of intermediate proteins, and fewer recent MRP proteins compared to the full proteome, with the difference (as illustrated by relative frequencies in Figure 4.6b) often reaching significance level.
Interestingly, when the GPCR pathway was compared to the full proteome only origin of its proteins was distributed differently (Figure 4.5c, MRP distribution was close to the full proteome trend (Figure 4.6b) However, it clearly stood out amongst other synaptic pathways through its recent diversification profile.

4.4.1.1 *GPCR signalling*

Based on evidence presented in this chapter I postulate that within the domain of proteomics the key factor in fine tuning of complex synaptic function is ongoing diversification of GPCR signalling pathways. Not only is it the only pathway studied here which gains a substantial number of proteins in more recent species (Figure 4.5) but also its earliest members continue positive selection well into recent branches of the root-human paths (Figure 4.7). Considering the variety of GPCR signalling cascades and their ubiquitous intertwining with molecular synaptic function (see the introductory section 1.2.1) they are a natural candidate for refinement of synaptic function. In Table 4.8, in the GPCR signalling row, even disregarding all olfactory receptors (*OR*s) there is a remarkable collection of early origin proteins with direct impact on synaptic plasticity (*PRKCB* and *PRKACB*) as well as proteins with confirmed effect on complex behaviour, such as *Oxitocin (OXT)*. Also, I used this pathway for a follow-up analysis to the hub effect introduced in section 4.2.5 and found further evidence for episodic diversifying selection events within the pathway graph.

4.4.2 *Selected proteins from pathways*

Following discussion of specific proteins identified through the extreme episodic selection profiles in Chapter 3 here I discuss a number of proteins relevant to the analysed pathways subjectively selected from Tables 4.8 and 4.9.

First, extending results about actin and microtubule cytoskeleton discussed in section 3.3.3; numerous *tubulins* appear in the 'conserved throughout' group in Table

4.9, only one *tubulin* each appears in the second and fourth column. However, many other protein members of the vescicle pathway appear in 'diversifying throughout' and 'diversifying recently' columns in the table. It allows me to hypothesise about the two stage development of the vescicle pathway, where basic skeleton appeared early and remained largely under negative selection, and only the finer regulatory agents diversified until recently.
The ribosomal proteins were another group of proteins which was hypothesised as a candidate for deep conservation in section 3.3.3. Here there is an abundance of their examples in 'conserved throughout' group, with a number of them in 'only early diversification' group too. Overall these results are consistent with the limited output of the enrichment tests, yet they provide more confidence in the magnitude of the deep conservation trend.

In the neurotransmitter pathway it is interesting to see *CAMK2* in the 'conserved throughout' column, while *PRKACB* appears in 'diversifying only recently' group. As introduced in section 1.2.3.3, *CAMK2* is a key element of the early LTP cascade - a relatively basic synaptic mechanism. According to the theory I tested through this chapter, a highly central, crucial protein is expected to be under mainly negative selection.
Then, *PRKACG*, an example of *PKA* family, is a regulatory protein involved in the same process as *CAMK2*, yet activated by the GPCR signalling cascade. *PKA* can phosphorylate AMPA receptor subunit to facilitate their insertion into the membrane yet this process is only regulatory and not necessary for the insertion (Esteban et al., 2003).

Further examples of *PKA* and *PKC* (another important regulatory kinase at the synapse, see section 1.2.3.3) can be observed in pre-NS column in Table 4.8. It means that despite early origin, predating nervous system, there is evidence for positive selection for them as recently as in the *H. sapiens* branch of the root - human path.
Further two interesting proteins featuring in Table 4.8 are *DNM2* and *OXT*. The first one is a key regulator of endocytosis but also plays role in cytoskeleton regulation at the developmental stages through interactions with microtubules and actin (Schafer et al., 2002; Raimondi et al., 2011). This multi-purpose synaptic protein is discussed in more detail in the following chapter.
Oxitocin is commonly described as a social bonding protein, as discussed in sec-

tion 3.3.2 social skills become vital in primates past the prosimian-simian breakpoint. However, a famous study into *Oxitocin's* role used prairie voles for observation (Ross et al., 2009). Also, it also has impact on non-social functions (Yang et al., 2013). Interestingly, it is implicated in theory of mind - a psychosocial construct encompassing social skill developed in childhood contributing to our understanding of other people's state of knowledge and emotions in order to predict their behaviour (Domes et al., 2007). Research into theory of mind outside human provides mixed results yet there is limited evidence for its existence in some of the primates closest to humans such as chimpanzees (Call and Tomasello, 2008; Heyes, 1998). Recent diversification of this protein observed here is fully consistent with these results.

Finally, the abundance of *Olfactory receptor proteins (ORs)* in Table 4.8, as well as in 'diversifying only recently' column in Table 4.9, illustrates the loss of complex function which occurred over the most recent divergence points leading to *H. sapiens*. As discussed in section 3.3.2, simians and more recently diverged primates gradually lost their olfactory skills as their behaviour patterns changed substantially.
In this particular case reduced importance of the function represented by that group of proteins results in reduced constraints on the sequence coding for them. In the modelling framework reduced constraint is not distinguishable from directional diversifying selection. Relaxation of selection constraints on olfactory proteins in Old World monkeys and human lineages has been discussed in literature already (Gilad et al., 2003, 2005), authors link this observation to environmental changes in a similar way as I previously mentioned in section 3.3.2.
In another study, Somel et al. (2013) observed a similar effect of relaxation for proteasome subunits. In this case the causal explanation remains unclear. The authors hypothesise about the environmental effect of an increased intake of dietary protein or a general trend of relaxation of constraint in human. It had been observed previously that small population lineages display weaker negative selection over the entire the genome, it might be the case that mutations considered to be slightly deleterious and which normally would have not been tolerated may achieve high frequency or even fixation due to drift (Kosiol et al., 2008).

In summary, the detailed results for pathways illustrate that different proteins within one pathway represent a wide range of selection profiles (including extreme opposites). Within each pathway there are key proteins providing the fundamental function of the pathway which are either under overwhelmingly negative selection

or completed their active diversification very early. The same pathway would also contain proteins which serve secondary, regulatory role; they are more likely to continue being under positive selection, or in extreme cases ('diversifying only recently' column) they remained under purifying selection until very recently, at which point they experienced an episodic diversification event.
This key finding can be considered an answer to the question about the link between a shared function and a shared selection timeline: this link is weak, and does not work as I had initially hypothesised. Instead, the protein's status within the pathway (or within the interactome) takes precedence over the shared function when determining evolutionary timeline.

Equally, in the context of synaptic function, interpretation of the pathway analysis results provides concrete evidence for how synapses capable of finely tuned plasticity evolved from early protosynapses. The basic function of propagating signal across the synaptic cleft was established early, yet layers of increasingly intricate regulation developed over time. In the early stages it was achieved through appearance of new proteins and active diversification of the existing ones, more recently, only active diversification of relative few proteins remained as the driving factor. And this mechanism of nervous system evolution continues until now as evidenced by the abundance of synaptic proteins diversifying in *H. sapiens* branch.

4.4.3 *Similarity of interactors and community members*

After average similarity of interacting proteins was not found to be different from average similarity of non-interacting proteins, community detection analysis continued the inquiry into the link between interactions and shared diversification history.
According to the analysis of similarities between proteins in identified communities only results for two communities supported the hypothesis about similarity of episodic selection profiles within interacting groups.
In the absence of functional annotation, such as in the form of pathways, communities derived from the interactome network can serve the role of partitioning a wide interaction network to functionally relevant groups of proteins, especially if they are anatomically limited such as in the case of the limitation (only PSP) I had imposed. Possibly integrating reliable data about anatomical patterns of protein expression could provide rationale for more complete partitioning of the full interactome and

communities could be identified in a similar way - providing a larger sample to test for the community-based evolutionary effects.

4.4.4 *Network centrality effects*

There are two alternative causal explanations for the effect of a relationship between most recent diversification, origin point, relative diversification window, and protein network centrality which I identified through the integration of network analysis and extracted measures of temporal profiles of proteins.

First, highly central proteins are important for the function of the neighbourhood complex so any non-synonymous mutations are likely to disrupt interaction paths between proteins; thus, they might be deleterious which is observed as strong purifying pressure. However, equally, reverse explanation is possible, and early appearing proteins under predominantly purifying selection might be able to form bindings with more proteins as they join the complex and as they evolve through their positive selection events.

The follow-up test of the hypothesis of propagation of positive selection provided unclear yet promising results which offer limited support to my theory. I found positive selection response in hub protein neighbours for hub appearance and for the oldest positive episodic selection event but not when all positive selection events were averaged together. Effects were generally clearer for the first layer of neighbours but the temporal aspect of the effect was hard to trace. Possibly, also the Yule-Simpson paradox can be blamed for unclear and sub-significant results when generalising over inherently heterogeneous groups such as proteins or branches here (Good and Mittal, 1987).

4.4.4.1 *Limitations*

First of all, interaction is only an indication of a suitability of two proteins to interact; it does not guarantee they are ever expressed in one tissue at the same time. This can occlude any true effects, as the postulated relationship between network features and evolutionary characteristics depends on interactions actually occurring and having meaningful consequences.
In a review of methods of acquiring interaction data von Mering et al. (2002) argue

that all high throughput experimental methods used for prediction of interactions reflect different phenomena which are measured. For example yeast two hybrid assays brings together two proteins and tests their interaction in an environment which is not natural for either of them. In comparison, tandem affinity purification captures protein complexes in their usual physiological setting yet the procedure itself (tagging, washing off) may tamper with the interactions.
In most proteome network analyses, including the current study, interactions from different experimental methods are all ultimately reduced to equivalent graph edges in an interaction network. Moreover, the authors claim that even 50% of interactions detected in high throughput studies might be irrelevant as they link functionally unrelated proteins known to appear in distant cellular locations (von Mering et al., 2002). We can also observe a bias towards interactions between proteins in certain processes; while very useful for studies focussing on a particular pathway, in a comprehensive study aggregating the entire proteome, such as here, this annotation bias results in the rate of false negatives and false positives varying between proteins.

In the specific case of my interaction network analysis there is an inherent trade-off between the signal-to-noise ratio and total coverage. I accepted a possible reduction in a clear signal by including all classes of interactions detected with different methods apart from computational predictions. Yet these broad inclusion criteria allowed for much wider analysis which is in line with the general goals of this research program, focussing on large sets of proteins and systematic effects.

Another potential limitation is temporal resolution of modelling data. First, among more distant species branch lengths are really long, even considering speed of observable evolutionary change. Second, although branch lengths between recent divergence points are relatively short, the taxa with missing annotation of orthologs become a problem, as one divergence point might be *supported* only by a single taxon. It is an entirely reasonable assumption that if a protein has an ortholog in both earlier and more recently diverged clades then it also has one in the intermediate one. Unfortunately some species, although fully sequenced, suffer from inadequate annotation of coding sequence.

4.4.5 *Interactions localisation*

Testing the effect of hubs on their neighbours returned promising yet unclear results. Possibly interacting partners only evolve synchronously at their respective binding locations; as suggested by Hakes et al. (2007), narrowly spatially isolated positive selection pressure would remain undetectable through the modelling framework I used to infer active diversifying pressure here (see section 2.1.7). I propose that only diversifying pressure directly affecting interaction would be the one that spreads across to interaction partners. I will try to address this hypothesis in the following chapter. I will look at a small complex of proteins with manually curated annotation data which includes binding domains. That approach requires a different modelling paradigm for selection pressure inference but I included it in my modelling pipeline hence these data are available (see section 2.1.7).

4.4.6 *Outstanding questions*

Compared to Chapter 3, this chapter offered more in-depth results due to incorporating other classes of data to the analysis: static protein-protein interactions, and functional pathway annotations. I revealed interesting effects of differences between post-synaptic interactome communities as well as systematic evolutionary timeline effects in the full human interactome collectively pointing to a link between the role of a protein in relation to other proteins and their profiles of episodic selection profiles. Further to that, pathways relevant to complex cognitive function show great diversity of their evolutionary characteristics between each other and compared to the full proteome background trends, lending further support to the early origin - recent diversification hypothesis already put forward in Chapter 3.

Considering promising yet unclear results of hub effect for both interactome and pathways I hypothesised that spatial profiles of selection pressure can shed more light on the issue of relationship between evolution and proteins' inter-dependency. Furthermore, if selection pressure acts on regions of proteins, then I can ask a more general question whether there are there any other sequence features apart from binding sites which could have systematically different evolutionary profiles to the rest of the protein. Results from spatial selection pressure modelling introduced in section 2.1.7 are available but have not been used yet. In the following chapter I will

present preliminary findings with these modelling data which aim to address the hypotheses mentioned above.

Table 4.8: Proteins with most recent (*H. sapiens* branch) evidence for episodic positive selection in selected pathways divided by their origin. See section 3.2.1.1 for further explanation of origin bins and section 3.2.3 for more information about using the measure of most recent positive diversification.

	Mammals	Vertebrates	NS	pre-NS
neurotransmitter	-	-	-	PRKCB, PRKACB
release cycle	-	-	PPFIA4	-
signaling GPCR	OR2T33, OR4C15, OR5D18, OR5K4, OR2T5, OR4A5, OR5K1	OXT, ARTN, TAS2R5, GFRA3, GFRA1, OR10H5, OR5B3, OR51T1, OR51F2, OR13D1, OR9A2, OR52N5, HCAR2, LPAR5, OR6C76, OR10J3, OR11L1, OR13G1	WNT8A, FGFR2, PDE4A, PDE10A, WNT6, GPR83, MCF2L, EMR2, RGS3, ARHGEF19, RGS10, WNT9B, NPY1R, FSHR, GPRC6A, NPSR1, WNT7B, GRM7, FGG, GRM5, NCAM1, PLCB4, PDE1A, SPTB	REEP6, ARHGEF11, ARHGEF7, DAB2IP, KRAS, PRKCB, PRKACB, RASAL2
translation	-	-	EIF3L	CARS
translation initiation	-	-	EIF3L	-

continued ...

. . . continued

	Mammals	Vertebrates	NS	pre-NS
vesicle	-	GOLGA2, GJD3	USE1, TRAPPC13, MASP1, FCHO2, VPS37C, DENND4A, CD163, SNX18, NAA38, ANKRD28, DENND1B, HPR, ANK1, DNAJC6, KLC1, SGIP1, SPTB	EPN2, AP4S1, COG4, CLINT1, PUM1, AP1S3, KIF6, GPS1, EPS15L1, EXOC6, DNM2, KIF3A, SCFD1, SYNJ1, VPS53
vesicle binding uptake	-	-	MASP1, CD163, HPR	-
vesicle membrane traffic	-	GOLGA2, GJD3	USE1, TRAPPC13, FCHO2, VPS37C, DENND4A, SNX18, NAA38, ANKRD28, DENND1B, ANK1, DNAJC6, KLC1, SGIP1, SPTB	EPN2, AP4S1, COG4, CLINT1, PUM1, AP1S3, KIF6, GPS1, EPS15L1, EXOC6, DNM2, KIF3A, SCFD1, SYNJ1, VPS53
axon guidance	-	ARTN, GFRA3, GFRA1	COL9A2, FGFR2, EPHB2, SCN3A, COL6A5, RELN, ANK1, CACNB4, FGG, NCAM1, SPTB	ARHGEF11, CACNA1D, ARHGEF7, DAB2IP, DNM2, ARHGAP35, KRAS, PRKACB, RASAL2
axon guidance RET	-	ARTN, GFRA3, GFRA1	FGFR2, FGG, NCAM1, SPTB	DAB2IP, KRAS, PRKACB, RASAL2

continued . . .

. . . continued

	Mammals	Vertebrates	NS	pre-NS
axon guidance EPH	-	-	EPHB2	ARHGEF7
NGF signalling	-	ARTN, GFRA3, GFRA1, CD19	FGFR2, MCF2L, ARHGEF19, FGG, NCAM1, PDE1A, SPTB	RAPGEF1, ARHGEF11, PIP5K1A, ARHGEF7, DAB2IP, DNM2, KRAS, PRKACB, RASAL2

Table 4.9: Genes with extreme diversification timelines (grouped according to four profiles described in section 3.2.4.1) in selected pathways

	Diversifying throughout	Diversifying only recently	Conserved throughout	Only early diversification
neurotransmitter	-	PRKACG	CAMK2A, GNAI2, GRIK4, KCNJ2, ADCY3, HRAS	GABRR3
releasecycle	APBA1, PPFIA4	-	GAD2	SLC1A3
signalingGPCR	PDE4C, SPTB, ARHGEF11, WNT9B, NCAM1, ARHGEF19, DGKI, LHCGR, RGS3, EMR2, FGG, OR6C76, OR5K4, HCAR2, GCGR	GLP1R, BDKRB1, PSMA5, C5AR2, RGS10, OR9A2, GPR35, OR9I1, OR13G1, OR5D18, OR2T8, OR2T10, OR2T5, OR2T33, OR2G2, OR11L1, OR4C15, OR5K4, OR2G3, OR5A2, NTSR2, OR5T2, OR52N4, OR52N5, OR1A1, OR13D1, OR52B6, OR51M1, PRKACG, OR4A5	PIK3CG, PSMD5, CAMK2A, ADCYAP1R1, PSMC6, PROKR2, PSMA7, GNAI2, PSME1, PRKCQ, PSMD10, SOS2, PRLHR, PPP3CA, MTNR1B, UBC, DUSP6, CRHR1, CRHBP, ADCY3, SHH, FRS2, HTR1F, PHB, PSMD1, HRAS	PIK3R3, RGS17, CX3CL1, PDYN, OR4D1, ADRA1A, OR7C1, OR11H1, TAC3, OR6C70, OR4F6, OR2B11, OR7A5, OR6C2, CXCL10, OR8G1, OR5M8, OR11G2, CSF2, ABHD6, OR5J2, TAS2R1, OR10H3, PDGFA, OR7A17, CSF2RA, IL3RA, OR4F21, OR6C68, OR6C1, OR4F16, OR6C65, OR8G5, OR5D16, OR4F3, OR52E6, OR4F29, OR11H2

continued ...

. . . continued

	One	Two	Three	Four
translation	-	RPS29	RPS13, RPL26L1, SEC61A1, GARS, RPL5, RPL23, ETF1, RPL10, RPL27, RPL37, RPL11, RPLP1, EIF1AX, RPL10L, RPL38, RPL10A, SSR2	RPS13, RPL26L1, GARS, RPL5, RPLP1, EIF1AX, RPL10A
translationinit	-	RPS29	RPS13, RPL26L1, RPL5, RPL23, RPL10, RPL27, RPL37, RPL11, RPLP1, EIF1AX, RPL10L, RPL38, RPL10A	RPS13, RPL26L1, RPL5, RPLP1, EIF1AX, RPL10A
vesicle	SPTB, SGIP1, TF, PPP6R1, CD163, NAA38, GOLGA2	CHMP2B, RAB5C, ACTR2, AP1S3, ALS2CL, TBC1D3, TUBB8	ARF5, ACTR3, RAB36, CHMP5, KIF16B, PAFAH1B1, VAMP4, TUBA3D, TUBB1, COPS5, GALNT2, RAB14, UBC, BICD1, VPS54, VAMP7, YWHAH, NEDD8, COPB1, COPS7B, GALNT1, TUBA3E, TRAPPC4, RAB11B, DENND5B, TMED3, BET1L, RALGAPB, TUBB4B, TUBB3	ARF5, RAB36, GJA8, TUBB4B
vesicleBinding Uptake	CD163	-	-	-

continued . . .

. . . continued

	One	Two	Three	Four
vesicleMembrane Traffic	SPTB, SGIP1, TF, PPP6R1, NAA38, GOLGA2	CHMP2B, RAB5C, ACTR2, AP1S3, ALS2CL, TBC1D3, TUBB8	ARF5, ACTR3, RAB36, CHMP5, KIF16B, PAFAH1B1, VAMP4, TUBA3D, TUBB1, COPS5, GALNT2, RAB14, UBC, BICD1, VPS54, VAMP7, YWHAH, NEDD8, COPB1, COPS7B, GALNT1, TUBA3E, TRAPPC4, RAB11B, DENND5B, TMED3, BET1L, RALGAPB, TUBB4B, TUBB3	ARF5, RAB36, GJA8, TUBB4B
axonguidance	SPTB, CACNA1I, COL9A2, ARHGEF11, SCN3A, EPHA5, NCAM1, FGG, COL6A5, SCN9A	CRMP1, PSMA5, ACTR2, DOCK1, ROBO1, KIAA1598, PRKACG, CSNK2B, TUBB8	PSMD5, ACTR3, CAMK2A, PSMC6, PSMA7, PSME1, PRKCQ, TUBA3D, PSMD10, TUBB1, SOS2, EPHA7, UBC, DUSP6, TUBA3E, DPYSL5, FRS2, PHB, TUBB4B, CDK5R1, PSMD1, HRAS, TUBB3	TREM2, PIK3R3, CSF2, PDGFA, TUBB4B, CSF2RA, IL3RA

continued . . .

. . . continued

	One	Two	Three	Four
axonguidanceRET	SPTB, NCAM1, FGG	PSMA5, PRKACG	PSMD5, CAMK2A, PSMC6, PSMA7, PSME1, PSMD10, UBC, DUSP6, FRS2, PHB, PSMD1, HRAS	PIK3R3, CSF2, PDGFA, CSF2RA, IL3RA
axonguidanceEPH	EPHA5	ACTR2	ACTR3, EPHA7, HRAS	-
NGFsignalling	SPTB, ARHGEF11, NCAM1, ARHGEF19, PIP5K1A, FGG	PSMA5, HDAC2, PRKACG	PRDM4, CASP2, PSMD5, CAMK2A, ADCYAP1R1, PSMC6, PSMA7, PSME1, PSMD10, SOS2, UBC, DUSP6, MAP2K5, ADCY3, FRS2, PHB, PSMD1, HRAS	PIK3R3, CSF2, PDGFA, CSF2RA, IL3RA

5

ARC COMPLEX SPATIAL ANALYSIS

The previous two chapters described global effects arising from comparing selection pressure modelling results across multiple proteins as well comparing groups of proteins to each other. These analyses informed general findings about the evolution of the entire human proteome as well as post-synaptic density.
However, in experimental science, there is much more focus on in-depth understanding of a handful of proteins at the time.
Analysis methods employed in Chapters 3, and 4 are not suitable for smaller sets of proteins as they depend on relatively weak and noisy global effects, with insufficiently understood underlying mechanistic principles. Also, previous chapters were limited to the temporal modelling data and did not use site-specific results of mixed-effects modelling framework primarily due to practical issues of systematic application and meaningful interpretation of them for large groups of proteins.
Therefore, in this chapter I proceed with analysis of proteins associated with *Activity-regulated cytoskeleton-associated protein (Arc)* using temporal as well as spatial data about selection pressure changes in an attempt to further investigate link between interactions and diversification timeline proposed in the previous chapter. Finally, I also provide preliminary results of an inquiry into evolutionary differences between structured and unstructured regions in proteins.

5.1 INTRODUCTION

As previously described in section 1.2.3.3 long-term synaptic plasticity involves multiple mechanisms of local assembly of protein complexes which are closely regulated in a fine network of multiple signalling cascades. Plasticity underpins high level cognitive processes such as memory formation (see section 1.2.3.3). *Arc*, also known as *Arg3.1* in literature, is one of the key regulatory elements for long-lasting plasticity, both LTP and LTD. Its impact on plasticity processes had been known already (Shepherd et al., 2006; Bramham et al., 2008, 2010), however, only recently its structure and interactions were studied in more detail (Zhang et al., 2015; Myrum et al., 2015) al-

lowing for reasoning about mechanistic principles of its molecular function.

The key high level cognitive correlate of *Arc* is memory formation as demonstrated in the knockout study (Plath et al., 2006). In terms of its clinical significance, *Arc* expression is affected in Fragile X syndrome and tuberous sclerosis (among many other proteins). Also, through its regulatory impact on *Amyloid precursor protein (APP)* and interaction with *Presenilin* it is involved in β-amyloid peptide production pathway and thus related to Alzheimer disease (Wu et al., 2011).

5.1.1 *Evolution of structure*

5.1.1.1 *Human protein structure*

Recently, two independent studies by Myrum et al. (2015) and Zhang et al. (2015) provided convergent evidence showing the mammalian *Arc* structure. In summary, the structure consists of two distinct domains - basic N-terminal domain and acidic C-terminal domain which further divides into N-lobe and C-lobe. The two domains are separated with a flexible linker/hinge region. Basic N-terminal domain promotes interaction with actin structures, and the protein is capable of reversible self-association which is hypothesised to create a scaffolding for binders, however, the functional significance of the phenomenon in the context of plasticity is still under investigation. Myrum et al. (2015) found that the only reported missense variant in human, located within N-lobe (V231G), does not affect structure in a significant way and has no measurable phenotypic effect.
Interactions occurring at the specific structural domains were also investigated by Zhang et al. (2015), a selection of them is further described in section 5.1.2 together with their biological significance.

5.1.1.2 *Emergence*

Bilobar C-terminal domain is homologous to retroviral *Gag* protein capsid domain (in viruses such as Human immunodeficiency virus and Rous sarcoma virus) (Campillos et al., 2006; Zhang et al., 2015). The earliest putative eukaryote orhtolog was identified in *D. melanogaster* (fruitfly), however, it lacks the N-terminal domain and contains a Zinc knuckle domain absent in mammalian *Arc* (Mattaliano et al., 2007). Orthology is not functional either, as fly protein is only implied in stress-induced behaviour,

not in memory formation. The next step of the gradual domestication and neofunctionalisation was observed in *d. rerio* (zebrafish). This is the most distant taxon in which N-terminal domain appears, however, the Zinc knuckle domain is only lost later (Mangiamele et al., 2010; Campillos et al., 2006). *X. tropicalis* (frog) *Arc* ortholog has a largely similar domain-level architecture as mammalian *Arc* despite only approximately 40% sequence identity; it is, however, expressed in the brain, and there is a widespread consensus that it is an ortholog of mammalian *Arc* (Mangiamele et al., 2010). From that point onward gradual changes occur through reptiles, birds and early mammals; in placental mammals the protein sequence and structure are remarkably conserved, the amino acid sequences of rodent and human Arc are identical.

Table 5.1: *Arc* spatial features, including regions involved in binding, nuclear function regulation domains, and high-level structural divsion. * it is not a region sufficient for binding, it is however necessary for it - I analyse together with all binding regions. See also Figure 5.1 for a visual representation.

Region type	Region name	AA location
binding region	binding PSEN1, UBE3A	1 - 154
binding region	binding SPTBN4	26 - 154
binding region	binding TFPT	67 - 396
binding region	binding SH3GL1, SH3GL3	89 - 100
binding region	binding RNF216	94 - 382
binding region	binding AP2A2, AP2B1, AP2S1, AP2M1	197
binding region	binding CACNG2, GRIN2B, GRIN2A, DLGAP1, CAMK2A, WASF1, IQSEC2	208 - 277
binding region*	necessary for binding DNM2	195 - 214
nuclear	nuclear retention domain	29 - 78
nuclear	nuclear export signal	121 - 154
nuclear	nuclear localization signal	331 - 335
structural	N terminal domain	25 - 134
disordered	central hinge	135 - 207
structural	N-lobe	208 - 277
structural	C-lobe	278 - 361

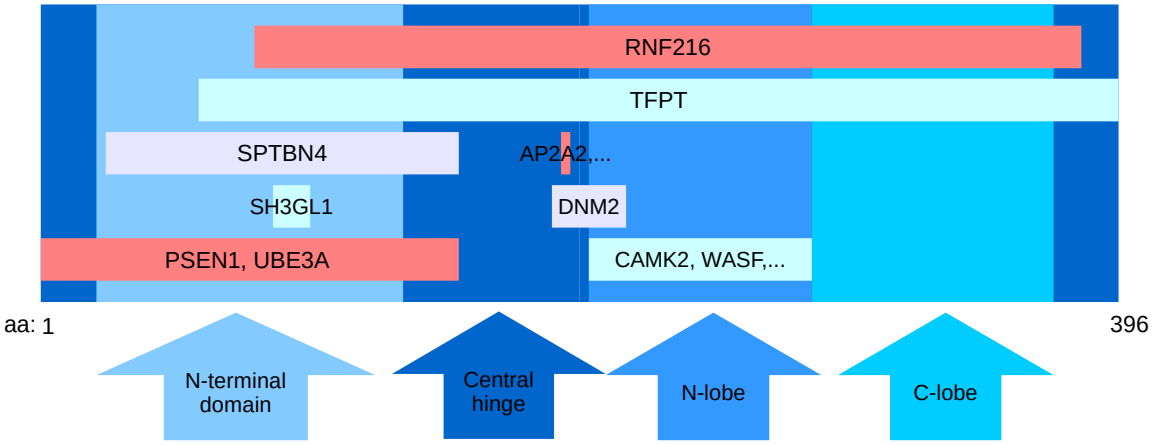


Figure 5.1: *Arc* spatial features based on Table 5.1

5.1.2 *Molecular role of the complex*

Arc is a multifunctional regulator at the synaptic density but also plays a role in regulating nuclear function. At the synapse, through its interactions with other proteins, it affects changes in dendritic spine size which translate to long-lasting LTP or LTD.

First, Arc was found to play a role in F-actin stabilisation contributing to spine growth in LTP, However, it does not directly interact with *Cofilin*, which in its phosphorylated state promotes cytoskeleton expansion. Instead, it interacts with *Drebin A* which competes for actin filament binding with *Cofilin*, thus indirectly regulating *Cofilin*-actin interaction (Nair et al., 2017). *Arc* also directly binds with *WASP1*, another actin-binding protein, part of *WAVE* complex which regulates cytoskeleton growth through interaction with *Arp2/3* complex (Zhang et al., 2015). Collectively, these interactions contribute to actin cytoskeleton elongation, branching, and stabilisation - essential for long-lasting synaptic strength increase.
Second, *Arc* interacts with *Endophilin-3* and *Dynamin-2* which are components of AMPA receptor endocytosis mechanisms. It was shown that *Arc* over-expression causes synaptic downscaling but the introduction of *Arc* without the *Endophilin* binding region does not have the same effect (Shepherd et al., 2006). Another molecular mechanism contributing to endocytic trafficking is *Arc* interaction with *Presenilin-1*

which, as part of γ-secretase complex generates β-amyloids from amyloid precursor protein (Wu et al., 2011).
Also, *Arc* molecular function at the synaptic density includes interaction with scaffold proteins such as *PSD95* and *GKAP*. The removal of *PSD95* by *Arc* disrupts the signalling cascade of BDNF receptor *TrkB* thus affecting synaptic scaling (Cao et al., 2013) (see section 1.2.2 for wider context).
Finally, although the main focus of this research is *Arc* local synaptic function, it is worth noting that the protein is also found in the nucleus, where its levels are regulated by synaptic activity in vivo, and it mediates *GluA1* transcription decrease, thus indirectly contributing to synaptic downscaling.

Table 5.1 and Figure 5.1 summarise the known interaction regions, and nuclear function regions based on a literature search. Not all interactions mentioned in this section can be tracked down to a specific region. The structural regions described in section 5.1.1.1 are also listed in the same table.

5.1.2.1 *Post-translational modifications*

Arc itself is regulated through multiple post-translational modification (PTM) sites. Particularly notable are the abundant ubiquitination sites playing a role in protein decay, degradation, and recycling. There are also multiple sumoylation sites which affect protein localisation. Notably, one of the sumoylation sites (at K268) overlaps with a ubiquitination site, however, contrary to expectations, sumoylation at this residue (blocking the possibility of ubiquitination) does not reduce *Arc* degradation rate. Overall, causal relationship between specific sumoylation sites and the functional role of Arc remains unclear and is a topic of ongoing research (Nair et al., 2017).
Finally, although they are the most abundant PTM type overall, there are relatively few phosphorylation sites indentified in *Arc*, and even fewer with evidence for functional relevance of a phosphorylation in vivo. A particularly notable site is S206 positioned at the border of the central linker and the N-lobe, within the *DNM2* binding region and close to the *AP2A2*, *AP2B1*, *AP2S1*, *AP2M1* binding region also in the linker. This site was found to be phosphorylated by *ERK* in vivo which led to increased cytosolic expression of the protein and is considered to be functionally relevant to conformational changes and interaction behaviour (Nikolaienko et al., in press).

5.1.3 *Objectives*

Arc is involved in multiple regulatory functions (actin cytoskeleton, endocytic trafficking, and nuclear regulation), for each of them a different subset of protein interactors is involved, however, it remains unclear which interactions are the most functionally relevant. Also, considering the high degree of conservation in mammalian *Arc* the outstanding question is how the functional complex evolved since that divergence point, and how it evolved prior to that moment; specifically around the time of *Arc* emergence because all *Arc* interactors have earlier origin points than *Arc* itself.

Here I aim to answer a selection of questions forming an extension of the ideas put forward in the previous chapter, as well as addressing wider question of evolutionary diversification differences between structured regions and unstructured regions:

1. First, does *Arc* appearance in an already established ecosystem of proteins trigger positive episodic selection in genes belonging to its complex in a similar way to how hub proteins affect their interactors (compare sections 4.2.5, and 4.3.3)?
2. Then, is this effect specifically localised to the experimentally derived binding sites between proteins?
3. Subsequently, I will use limited domain annotation of selected proteins and complete proteome-wide PTM site annotation to test whether selection pressure acts differently in these specific regions of proteins compared to the rest of the sequence.
4. Finally, I will explore the possibility of using the evolutionary profiling of putative members of *Arc* complex to help differentiate true complex members from false positives.

5.2 RESULTS

5.2.1 *Protein set inclusion criteria*

I studied a set of 138 proteins: 27 interactors of *Arc* from literature; 111 proteins identified as significantly differentially copurified with *Arc* (O. Nikolaienko & C. R.

Bramham, personal communication), where one was also in the interactors list; and finally, *Arc* itself (see Table 5.2 for the full list).

5.2.1.1 *Interactors*

Confirmed interactors were identified in publicly available interaction databases and limited to low-throughput experiments on rodent (*M. musculus* and *R. norvegicus*) and human proteins. For 16 experimentally confirmed interactors data about the regions in these proteins playing a role in binding were available; for 19 of them data about *Arc* regions were available, however, in the case of *Dynamin-2* it is a necessary but not sufficient region for interaction (Nikolaienko & Bramham, personal communication). Overall, regions in both proteins are known for 14 interactors. Domains in interactors are listed in Table 5.3, domains in *Arc* are listed in Table 5.1.

5.2.1.2 *Copurification*

The second group of proteins, which consists of putative members of *Arc* complex, were identified in a experiment using differential copurification with *Arc* (Nikolaienko & Bramham, personal communication), Kimple et al. (2001) provides a review of the experimental methodology. The significance cutoff was set at $p < 0.05$ and proteins which were significantly differentially copurified with *Arc* but are known to be common contaminants in such expreriments were eliminated - these included various keratins (hair component).

5.2.2 *Spatial modelling data*

Full results for three modelling methods (see section 2.1.7) were used in this chapter. A summary of the results from each of the methods for all *Arc* complex genes is presented in Supplementary Table A.7 for reference. Since Chapters 3 and 4 focussed on temporal aspect of selection pressure through aBSREL model, here, I focus on the outputs of the spatial modelling paradigms (FEL and MEME).

5.2.3 *MEME and FEL comparison*

In section 2.1.7 I described two methods of site-specific inference of evidence for positive selection pressure (FEL and MEME), and calculated results of these modelling ap-

Table 5.2: Proteins in *Arc* complex from co-purification experiment and interactors sourced from literature; * protein appears in both lists, †spatial annotation of a region relevant for binding is available (see Table 5.3).

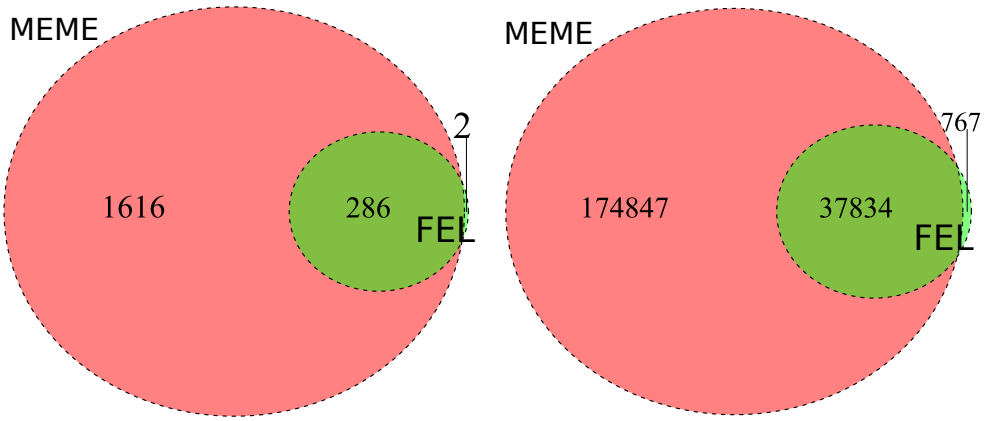
Experiment			Interactors
VDAC3	CYB5B	TM9SF2	AP2S1
SLC2A1	CD99	SGPL1	CAMK2A
COG4	ARFGAP1	SEMA4C	MAP2
TM9SF3	LAMTOR3	CCNDBP1	DNM2†
SLC1A5	HSDL1	TVP23C	RNF216*†
LPHN2	FXYD6	OSTC	WASF1
HPCAL1	ARL1	DNAJC7	CAMK2B
CCDC47	RTN1	SLC25A10	PSEN1†
TNPO3	RTN3	CRELD1	UBE3A
TNPO2	PTRH2	SRPR	AP2B1
ERGIC1	MRPL15	TUBA1A	CREBBP
M6PR	PPP3CA	IKBIP	SH3GL3†
VEZT	STT3A	CEP44	PML
HM13	ASH2L	ITFG3	NOTCH1
NOL4	C1orf35	TVP23B	SH3GL1†
SCAMP3	CNNM4	METTL7A	DLG4
SEC62	RAP1GDS1	RBBP4	KAT5
ARMCX3	SLC25A19	UBE2Q1	CACNG2†
VAPA	SEC24D	TADA3	AP2M1
SEC63	RNF170	USP7	GRIN2A
SLC38A1	PTGER2	LETM1	AP2A2
ABHD12	THEM6	PRKAG1	DLGAP1
PRKAG2	EFNB2	VDAC2	SPTBN4
GNAI2	RPLP1	NOL4L	GRIN2B
GNAO1	ERGIC3	RXRA	TFPT
RNF216*†	KRBA1	NAP1L5	DBN1
RNGTT	GNA13	YIPF6	IQSEC2
RAB18	TMEM35	USMG5	
AKAP8	CERS2	TUBB6	
SDHA	RHOT1	PPP2R2A	
C1QBP	SLC35E1	AATF	
NOTCH3	CNNM2	SCAF8	
COX15	SRPRB	NME1	
PPM1A	ZMYM2	ATP5J2	
MAVS	C10orf88	C5orf51	
GOLPH3	DNAJC14	RXRB	
SLC39A14	EIF2B2	SMIM1	

Table 5.3: Arc interactors binding regions.

Protein	AA location	Publication
SH3GL1	172 - 368	Chowdhury et al., 2006
SH3GL3	172 - 347	Chowdhury et al., 2006
DNM2	503 - 871	Chowdhury et al., 2006
PSEN1	1 - 49	Wu et al., 2011
RNF216	201 - 470	Mabb et al., 2014
CACNG2	221 - 247	Zhang et al., 2015
IQSEC2	1329 - 1385	Zhang et al., 2015
WASF1	301 - 343	Zhang et al., 2015
CAMK2A	278 - 329	Zhang et al., 2015
DLGAP1	436 - 452	Zhang et al., 2015
GRIN2A	1169 - 1174	Zhang et al., 2015
GRIN2B	1385 - 1390	Zhang et al., 2015
AP2A2	1 - 621	DaSilva et al., 2016
AP2B1	1 - 591	DaSilva et al., 2016
AP2S1	1 - 142	DaSilva et al., 2016
AP2M1	1 - 435	DaSilva et al., 2016

proaches for all proteins. MEME is discussed as characterised by far superior power in positive site detection (Murrell et al., 2012). Although a systematic simulation based test of this claim is beyond the scope of this thesis, here I visualise the striking difference between two approaches when applied to real data - an overlap between sites identified as positive with the same significance threshold ($p < 0.05$) for *Arc* complex proteins in Figure 5.2a, and full proteome in Figure 5.2b. It is clear that positive sides detected by FEL are practically a subset of positive sites detected by MEME.

Further to Figure 5.2 I investigated 767 significantly positive sites according to FEL model which did not reach significance threshold with MEME approach. The distribution of p-values for these sites is tightly clustered in the marginally sub-significant region (see Figure 5.3) which means they were modelled as positively selected but narrowly missed the arbitrary threshold of significance; relaxation of the p-value threshold for MEME would have included them as positives.
In summary, based on the observations above I agree with MEME's superior recall of sites under positive selection. Therefore throughout the rest of the chapter I use MEME as the source of site-specific selection pressure modelling estimates.



(a) Arc complex subset (b) Full human proteome

Figure 5.2: Overlaps of positive sites identified by two methods (FEL and MEME), note a very small number of sites in FEL set but outside MEME set.

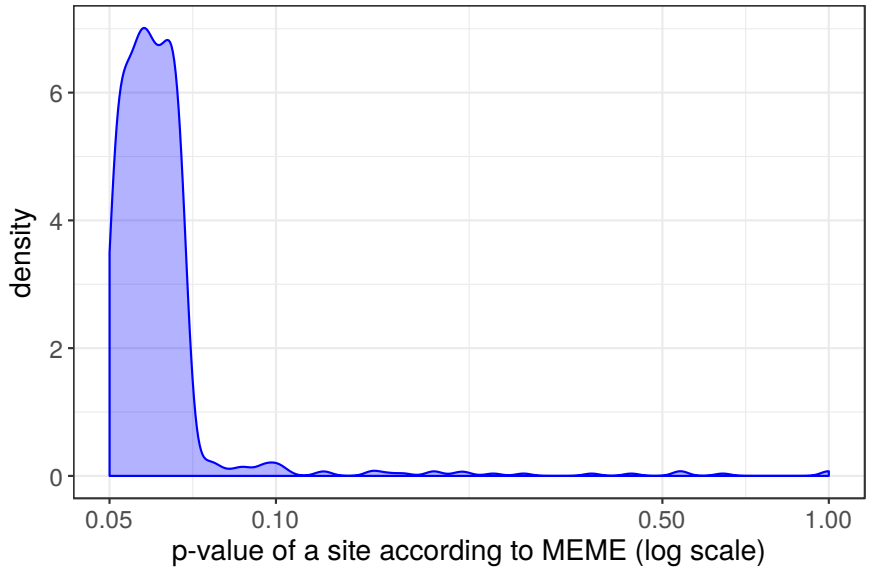


Figure 5.3: Distribution of MEME p-values for sites significant in FEL but missed by MEME. Note the majority of sites occupy the sub-significant level between 0.05 and 0.1.

5.2.4 *Sequence annotation data*

5.2.4.1 *Domains*

Partial domain annotation was initially sourced from Uniprot (Bateman et al., 2015) but upon inspection it was discovered to be unreliable with no clear definition of what constitutes a domain in the database; there was no systematic way of distinguishing sequence repeats, structural units, and binding domains without extensive

manual curation. After limited manual screening all domains annotated as *binding*, *interacting*, or *necessary for binding/interaction* were marked as binding regions.

Detailed, manually curated annotation of *Arc* itself and *Arc* binding domains in its interactors were compiled with help from O. Nikolaienko & C. R. Bramham (personal communication), full functional and binding domain annotation as well as structure was only used for *Arc*. These sequence features are listed in Table 5.1 for *Arc* and Table 5.3 for its interactors.

Table 5.4: Frequencies of PTMs used in protein annotation

Modification type	Count in full proteome	Count in *Arc* complex
Acetyl	20299	243
Methyl	14990	163
O-GalNAc	2016	3
O-GlcNAc	412	3
Phosphoryl	222298	2044
Sumoyl	6671	71
Ubiquitin	39620	556

5.2.4.2 *PTM*

Phosphosite (Hornbeck et al., 2004, 2015), the leading source of high quality annotation of all classes of PTM, was selected as the source of locations here. Counts of different classes of modifications in the full proteome and in *Arc* complex are listed in Table 5.4.

5.2.5 *Temporal landscape at full protein level*

The nalysis of aBSREL modelling output was not the main point of this chapter as I devoted two previous chapters to a systematic study of temporal effects of episodic selection pressure, however, here I visualised members of *Arc* complex in a similar way to Figure 3.7. Confirmed interactors are plotted separately in Figure 5.5, it is clear that they have much less evidence for recent positive selection (with the exception of *DNM2*).

In the confirmed interactor set there is very limited evidence for any positive episodic selection past the level of the split between the rodent lineage and the primate lineage

(second Euarchontogriles from the top in the figure) and the last broad sweep across most of the proteins (although sub-significant evidence for some) occurs at the level of Eutheria (Placental mammals).

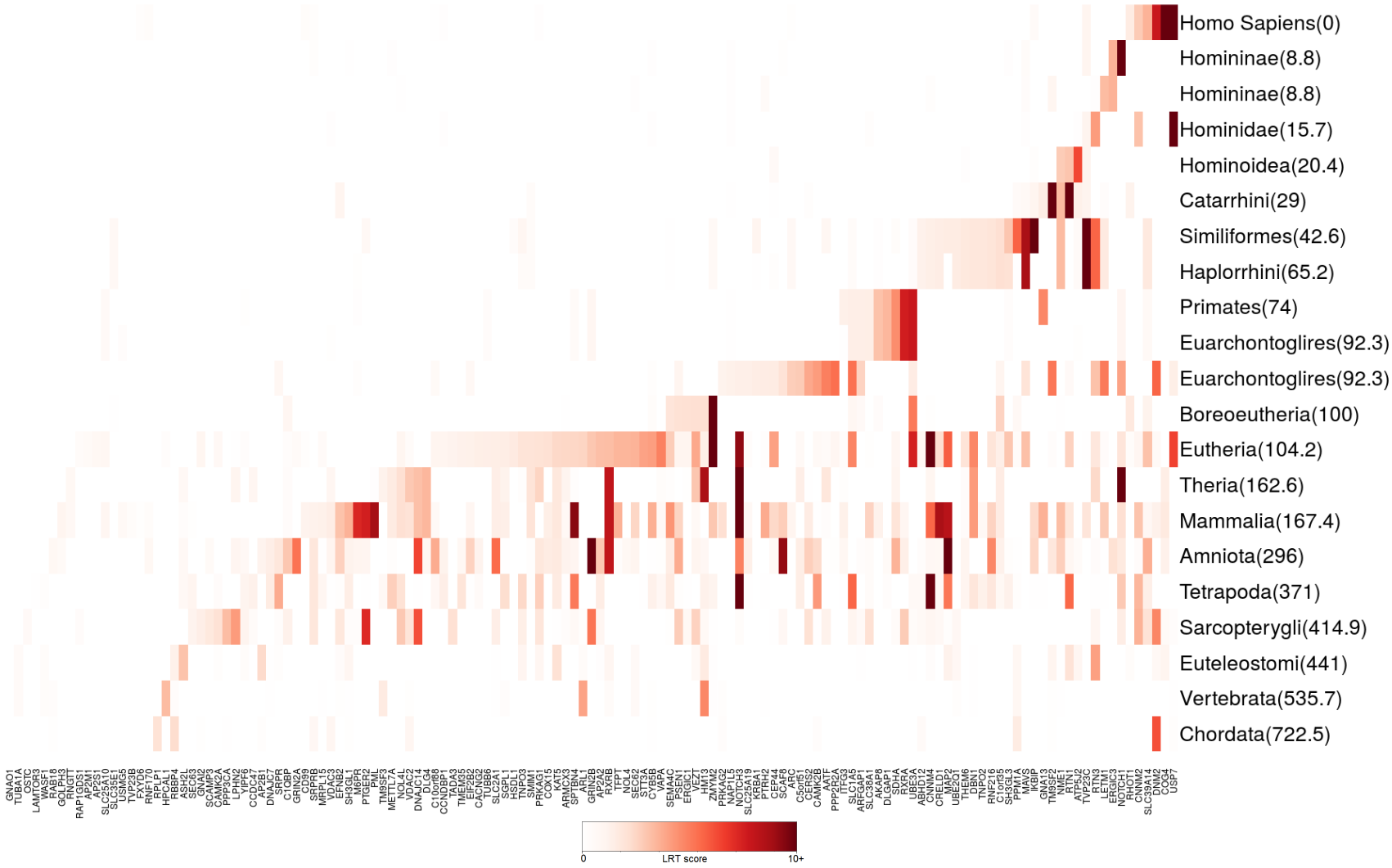


Figure 5.4: Most recent positive diversification for all members *Arc* complex. Compare to the general trend observed in the full proteome in Figure 3.7.

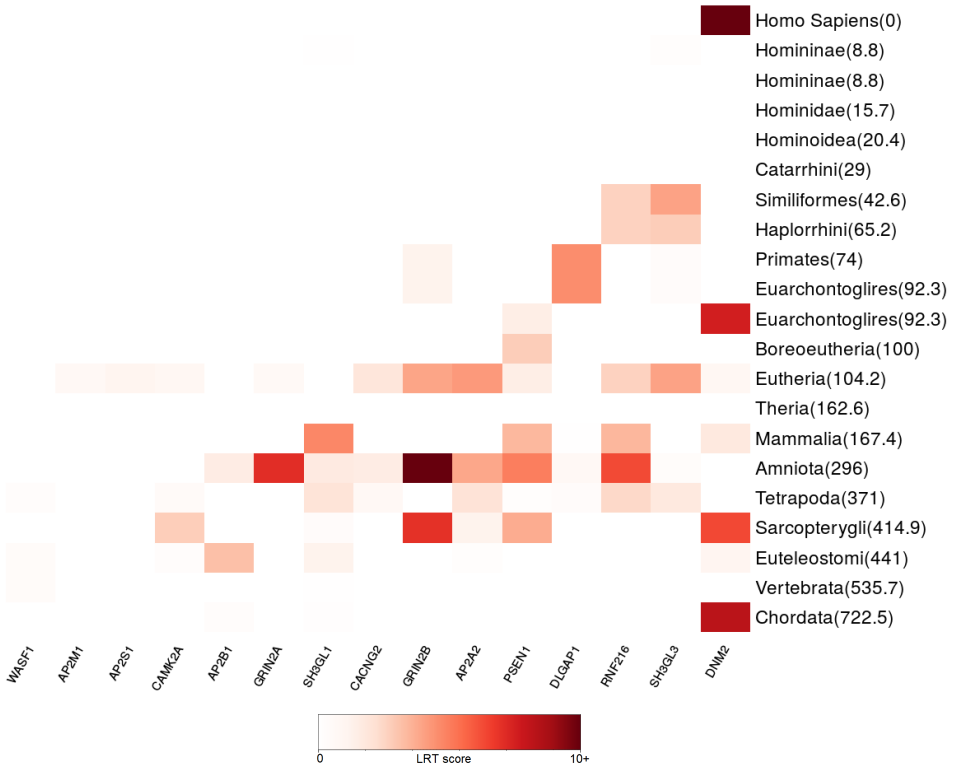


Figure 5.5: Most recent positive diversification for confirmed interactors of *Arc*. Note limited evidence for recent diversification.

5.2.6 *Spatial effect on domain level*

First, I tested the hypothesis about the difference between aggregated levels of positively selected sites inside regions involved in binding and the remaining amino acids.

I restricted tested regions to ones which cover at least 1% of the length of the interactor in order to ensure a sufficient sample size (in this case sample size is the number of amino acids inside the domain). I also eliminated regions defined too broadly, such as in all adaptor-related protein complex subunits (*AP2A2*, *AP2B1*, *AP2S1*, *AP2M1*) as they cover 66% – 100% of the protein length. Overall, 9 binding regions remained and results for them are presented in Table 5.5.

Since *Arc* has only 4 sites with significant evidence for local positive selection, which is a substantially lower frequency of positive sites (0.0101) than average in the full set (0.0645). A systematic comparison of corresponding regions involved in bind-

Table 5.5: Spatial selection pressure differences in binding domains of Arc interactors; comparison is a difference and ratio of frequencies of positive sites inside and outside the domain; p-values of a Chi-square test of difference between distribution of positive/negative sites inside vs. outside the domain

	Domain	Positive sites frequency		Comparison		
Protein	length	in domain	outside	difference	ratio	p-value
DLGAP1	17	0.588	0.084	0.495	6.972	<0.001
PSEN1	49	0.306	0.079	0.203	3.878	<0.001
WASF1	43	0.163	0.039	0.114	4.200	0.003
CAMK2A	52	0.135	0.061	0.066	2.206	0.130
DNM2	369	0.095	0.050	0.045	1.901	0.023
SH3GL1	197	0.096	0.082	0.007	1.178	0.792
SH3GL3	176	0.074	0.064	0.005	1.148	0.908
RNF216	270	0.067	0.096	-0.020	0.697	0.245
CACNG2	27	0.000	0.111	-0.102	0.000	1.000

ing with its interactors would not be informative in this case.

Using binding region annotation for all 138 *Arc* complex proteins I found that for most domains the difference between the frequencies of positive sites inside and outside the domains is negative but with a long thin tail of positive values. A comparison of fitted density plots for the frequency of positive sites in and outside the domain reveals how these two distributions vary despite exhibiting a similar mean value (see Figure 5.6). The distribution of differences between frequencies is visualised in the histogram in Figure 5.7. The binding regions of *PSEN1*, *WASF1*, *CAMK2A*, *DNM2* listed in Table 5.5, with a higher frequency of positively selected sites inside the domain compared to the outside, are unusual compared to the general trend as they occupy the tail of the difference distribution in Figure 5.7; they will be discussed further in this chapter. The effect for *DLGAP1*, although the largest in size and highly significant, needs to be evaluated in the context of how short this binding region is (only 17 amino acids).

5.2.7 *Spatio-temporal effect at domain level*

Part of MEME modelling output is an indication of specific combinations of sites and branches where positive selection occurred (see section 2.1.7). These spatio-temporal

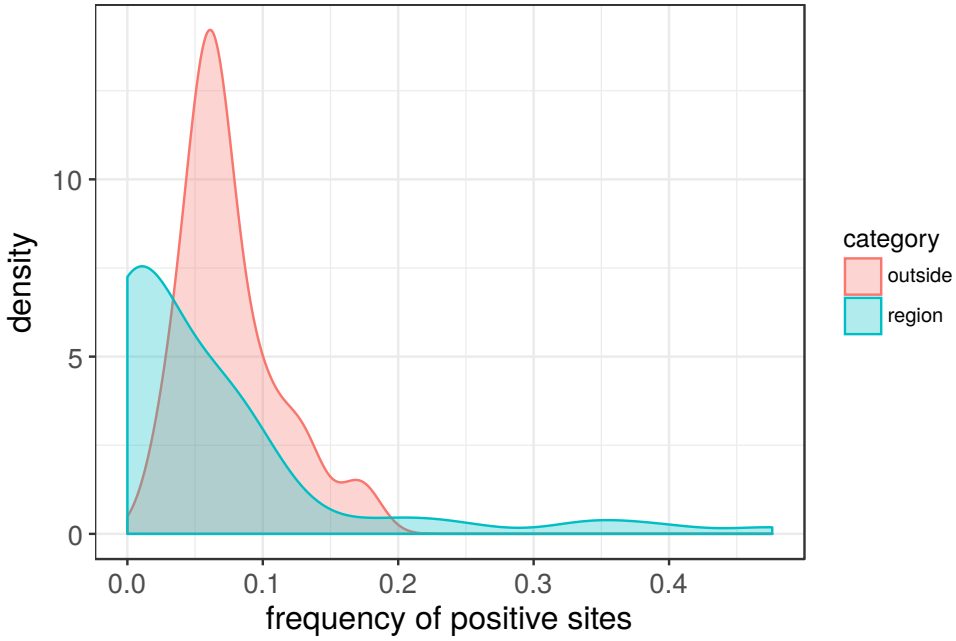


Figure 5.6: Density plots of frequencies of positive sites in and outside binding regions

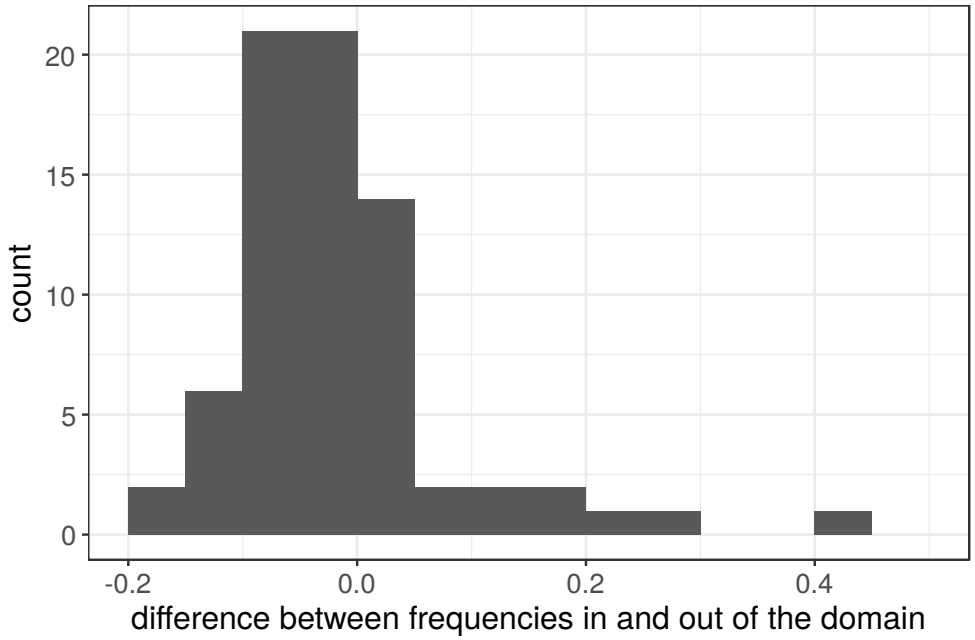


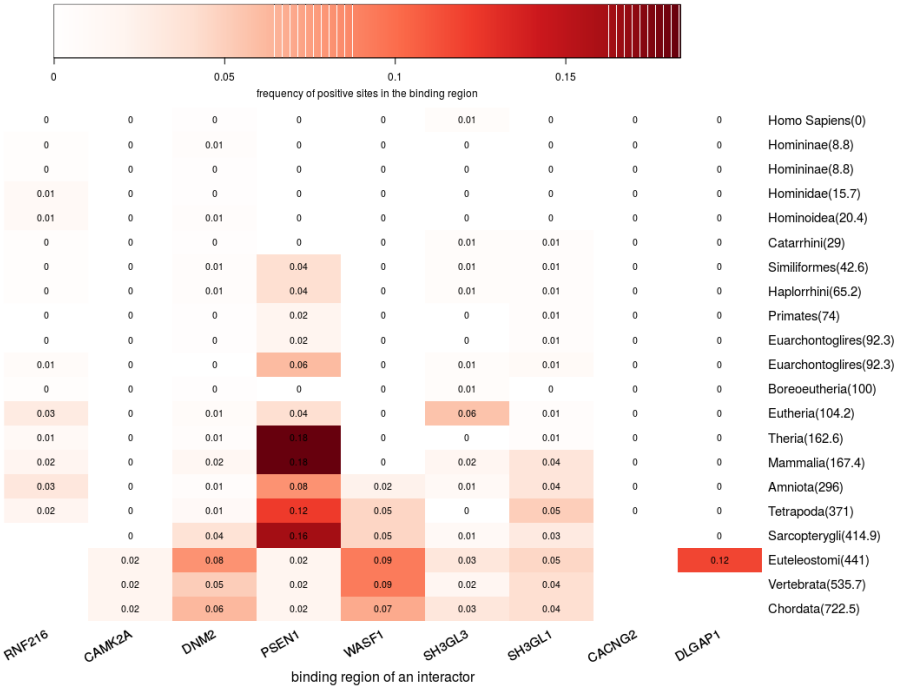
Figure 5.7: Histogram of difference between positive site frequencies between in and outside binding region

estimates in the form of EBF (ratio of likelihoods integrated over parameters) allowed for an analysis of effects on the intersection of temporal and spatial aspect of selection pressure events. I decided on a low threshold of $EBF > 3$ (Kass et al., 1995) as I did not aim to make any definite conclusions for the individual sites, instead, I looked

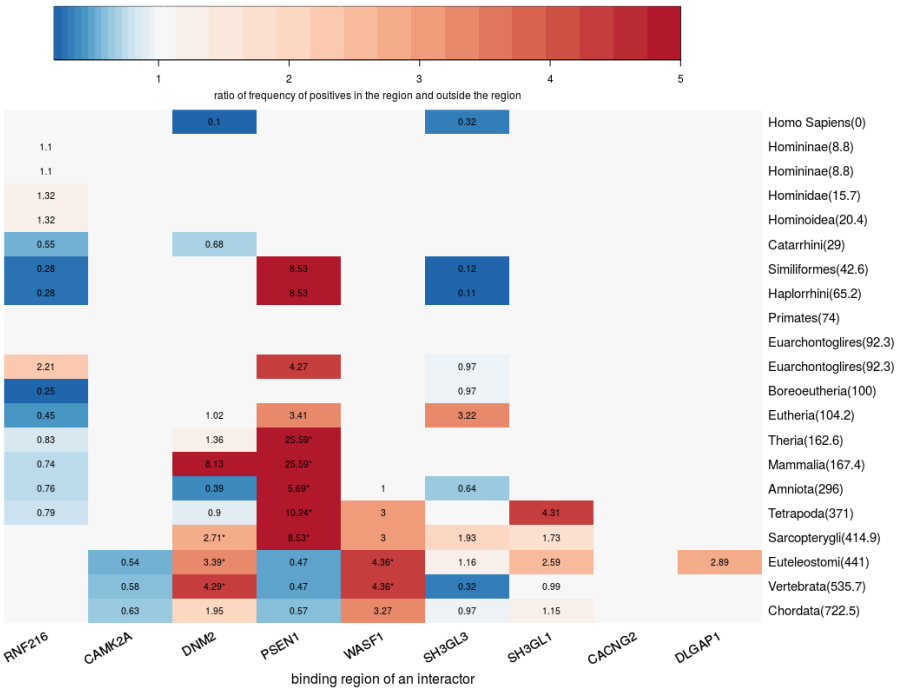
for broader patterns spanning entire domains.

In order to visualise the temporal profiles of binding regions, first, I calculated frequency of sites representing positive selection pressure (according to the criterion above) inside the binding region for each branch (see Figure 5.8a).

Then, I calculated a corresponding frequency for protein sequence outside the binding region, the ratio of two frequencies is plotted in the heatmap in Figure 5.8b. It allows for the detection of differences between the binding region and the rest of the protein as an equally high/low frequency of positive sites across the entire protein is not as informative for answering my research question. Cells with missing values in Figure 5.8b are ones for which there were no positive sites in the binding region, or no positive sites in the rest of the sequence, and statistical significance indication follows the same rationale as in section 5.2.6.

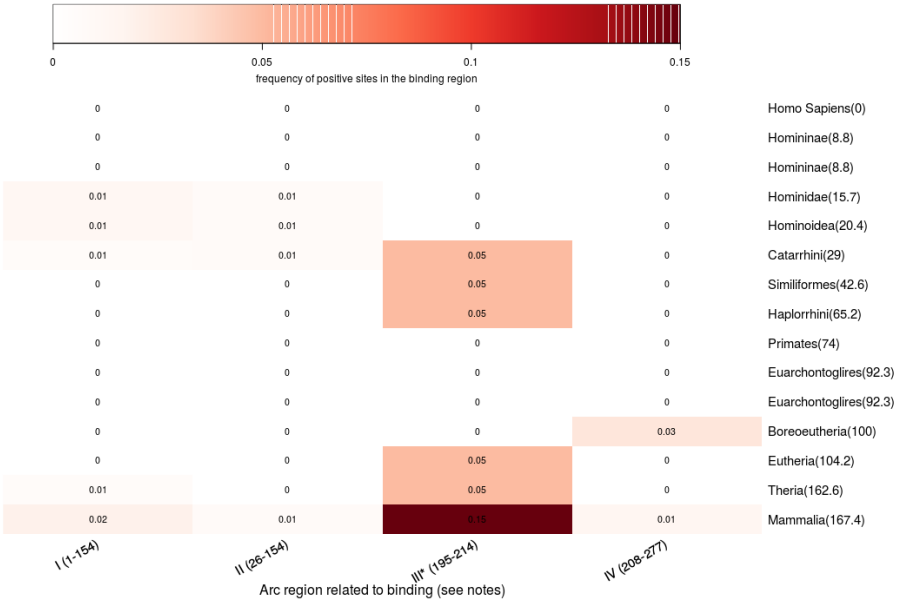


(a) Frequencies of positive sites in binding regions

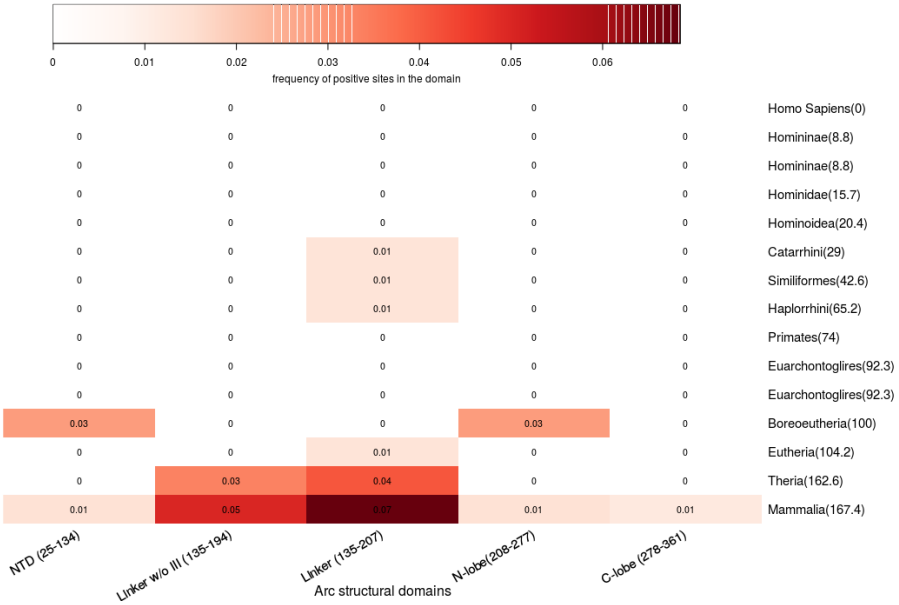


(b) Ratio of frequency of positive sites in binding regions over frequency outside the region

Figure 5.8: Temporal signature of domain-specific diversifying pressure of *Arc* interactors, see Table 5.3 for more details about binding domains. See section 5.2.7 for details about generating numbers in heatmaps above. Cells with missing data in 5.8b occur whenever there where no positive sites either inside the domain or outside of it.



(a) Binding regions



(b) Structural features

Figure 5.9: Temporal signature of domain-specific diversifying pressure of *Arc* structural domains and binding regions.
In the top heatmap Region I is involved in binding *PSEN1* and *UBE3A*, region II - binding *SPTBN4*, region III is a region which is necessary for *DNM2* binding but not sufficient, region IV is involved in binding *CACNG2*, *GRIN2B*, *GRIN2A*, *DLGAP1*, *CAMK2A*, *WASF1*, and *IQSEC2*; also see Table 5.1.
In the bottom heatmap a separate column is added for the portion of the linker region not involved in binding *DNM2*.

In the case of *Arc* (see Figure 5.9), similar to the spatial analysis in previous section, there are very few sites under positive selection at any of the branches. They tend to be located in the linker region, which largely overlaps with binding region III - the region necessary for *DNM2* binding, hence I added a separate column for the linker region excluding the portion which is the *DNM2* binding region. Overall, for these 3 regions (practically for one broad region) only the two earliest branches are affected to a degree which gives confidence in the results.
As a result, testing the timeline similarities between binding regions in interacting proteins (e.g. *Arc* region binding to *PSEN1* with *PSEN1* region binding to *Arc*) would not be informative hence the hypothesis about direct profile similarity between binding domains cannot be tested in this case.

5.2.8 *PTM effects*

Motivated by the question of the relationship between selection pressure events and functional sites at amino-acid locations I tested the frequencies of positive sites among all PTM sites (and among all classes of PTM sites separately) in comparison to the frequency of positive sites in non-PTM sites (the background frequency). The results are summarised in Table 5.6; overall, the frequency of positive sites in PTM sites was found to be higher than in non-PTM sites.
However, the direction of the effect varies between different classes of PTMs, e.g. sumoylation sites are significantly less likely to be under positive selection pressure whereas phosphorylation sites are significantly more likely to be under positive selection pressure; phosphorylation sites were the most abundant overall, hence the direction of the aggregated effect. Interestingly, when similar tests are repeated for *Arc* complex proteins (see Table 5.7) the overall effect for all PTMs remains similar (yet non-significant) while the direction of the effect for different classes of PTM varies compared to the full proteome (also non-significant).
The effect of PTM presence in human sequence on positive selection at a site was then tested within generalised linear mixed-effects model framework (McCulloch, 1997) on a sample of 10% proteins. Each site constituted a single datapoint, significant positive selection pressure was the response variable (binary, distributed according to Bernoulli distribution), PTM presence at the site was the fixed effect and the protein from which the site came was the random effect. The model was fitted with the numerical method which is generally discussed as the most precise (Monte Carlo Like-

lihood Estimation) and implemented within `glmm` package (Knudson, 2016) which allowed for attaching significance values to fitted coefficients.

The fixed effect coefficient for PTM presence in human sequence was non-significant ($p = 0.15$); the random effect of the protein was significant ($p = 0.013$). Results suggest that at any site positive selection is not affected by PTM presence when we account for natural variability between proteins.

Table 5.6: PTM positive selection pressure frequencies in the full human proteome (for significant change +/- sign indicates direction of change compared to non-PTM sites), p-values of Chi-square test for difference compared to non-PTM sites distribution corrected for multiple comparisons.

Type of PTM	total sites	positive sites	positive frequency (direction of change)	corrected p-value
Non-PTM sites	9778904	673837	0.0689	-
All PTMs	293989	24074	0.0819(+)	<0.001
Acetyl	19985	1321	0.0661	0.121
Methyl	15291	1056	0.0691	0.953
O-GalNAc	1949	154	0.0790	0.086
O-GlcNAc	400	31	0.0775	0.562
Phosphoryl	211107	18813	0.0891 (+)	<0.001
Sumoyl	6658	382	0.0574 (-)	<0.001
Ubiquitin	38599	2317	0.0600 (-)	<0.001

Table 5.7: PTM positive selection pressure frequencies in the full human proteome, p-values of Chi-square test for difference compared to non-PTM sites distribution corrected for multiple comparisons.

Type of PTM	total sites	positive sites	positive frequency	corrected p-value
Non-PTM sites	74963	5066	0.0676	-
All PTMs	3083	229	0.0743	0.158
Acetyl	243	12	0.0494	0.317
Methyl	163	17	0.1043	0.088
Phosphoryl	2044	160	0.0783	0.064
Sumoyl	71	9	0.1268	0.080
Ubiquitin	556	31	0.0558	0.307

Also, in Figure 5.10 I visualised the distribution of frequency of positive sites among PTMs and non-PTM sites summarised by protein; the effect is similar to the one observed in Figure 5.6. For most proteins functional regions, here PTMs, have lower

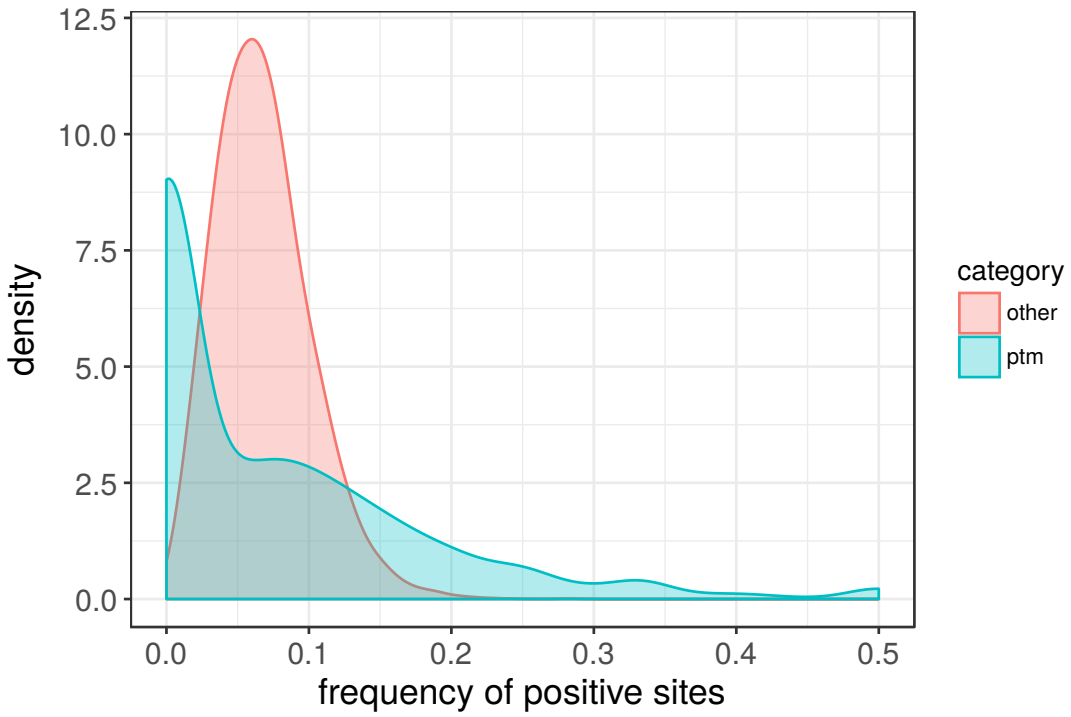


Figure 5.10: Density of frequencies of positive sites among PTMs compared to other sites in full proteome summarised by protein.
.

chance of being under diversifying selection pressure but long tail of the distribution signals reverse trend for some proteins. It illustrates how a systematic comparison, such as the one presented in Tables 5.6 and 5.7, does not fully reflect the inter-protein variability in the direction of the effect.

5.2.9 *Possible pipeline modifications*

An argument can be made for limiting the taxonomic diversity of orthologs used for spatial modelling analysis to the point of emergence of *Arc*. This modification has the potential to reduce gaps in MSA which is supposed to improve the accuracy of downstream selection pressure inference (Talavera and Castresana, 2007). Also, since *Arc* is the central protein of the complex we may only be interested in functionalisation of other proteins in the context of *Arc* being present. However, my systematic method of ortholog acquisition only resulted in the *Arc's* most distant ortholog being available in *X. tropicalis*, which is in fact quite recent considering an earlier *D. rerio* ortholog of *Arc* already consists of all main structural elements of the protein as we know it in rodents or human. Hence, in its temporal and spatio-temporal aspects

the modified modelling analysis would in fact miss the period of gradual process of *Arc*'s functionalisation which I initially hypothesised would be most interesting for the formation and refinement of the entire complex's function. For the sake of completeness, I briefly compared the results for the modified pipeline. In the ortholog acquisition step I limited taxonomic diversity to 51 out of 68 species available in Ensembl Compara setting a cut-off based on the most distant (measured by tree divergenece points) ortholog of *Arc* identified within constrains of my workflow - *X. tropicalis* - which maps to origin point 17 in table 3.1.

The remaining steps of the modelling workflow remained the same. For each input protein the output of this step was a set of unaligned ortholog transcripts; the numbers of orthologs for each gene are reported in Supplementary Table A.6.

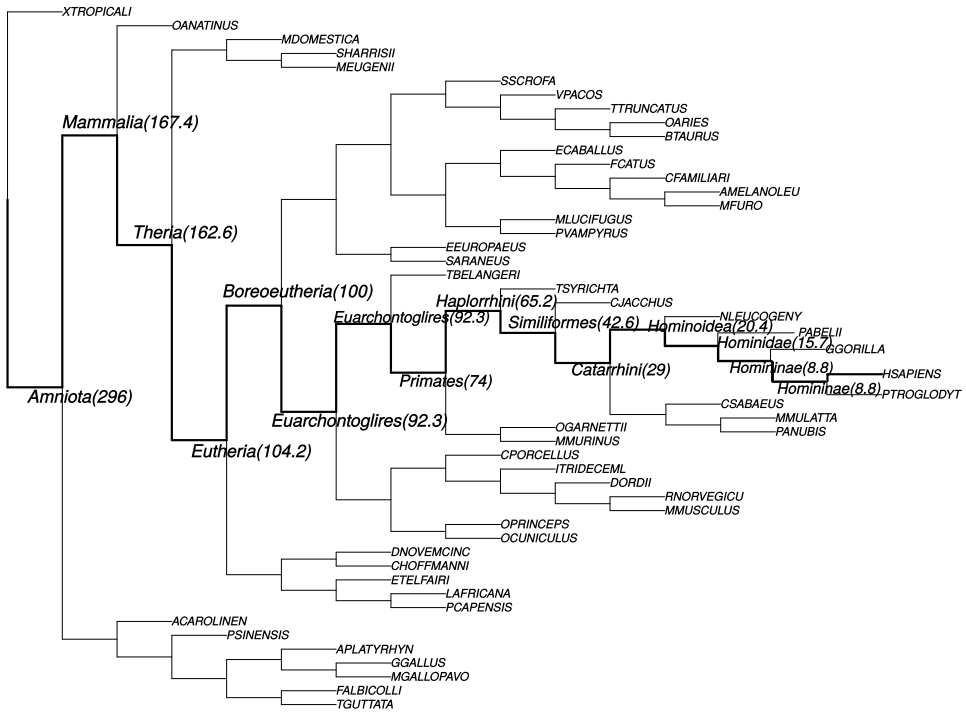


Figure 5.11: Reduced phylogenetic tree used for an alternative *Arc* complex analysis. Internal nodes' dating in parentheses expressed in *mya* according to Ensembl Compara; highlighted branches form a path from the root to *H. sapiens* leaf.

The spatial effects discussed earlier in section 5.2.6 remained, yet with a smaller magnitude of the effect (see Table 5.8). This smaller magnitude of the spatial effect is fully understandable considering the temporal aspect of site-specific selection pressure at the protein level (see Figure 5.5). Further to that, the domains of interest had evidence for positive sites at early branches, preceding appearance of *Arc* defined as origin point 17 (see Figure 5.8).

Table 5.8: Spatial selection pressure differences in binding domains of Arc interactors (optional reduced tree analysis); order retained from Table 5.5 to facilitate comparison, see section 5.2.7 for description of procedure.

	Domain	Positive sites frequency		Comparison		
Protein	length	in domain	outside	difference	ratio	p-value
DLGAP1	17	0.353	0.070	0.278	5.057	0.001
PSEN1	49	0.163	0.081	0.073	2.007	0.153
WASF1	43	0.047	0.045	0.002	1.043	1.000
CAMK2A	52	0.192	0.068	0.111	2.825	0.013
DNM2	369	0.095	0.044	0.029	2.160	0.008
SH3GL1	197	0.102	0.070	0.015	1.447	0.429
SH3GL3	176	0.051	0.082	-0.015	0.625	0.389
RNF216	270	0.041	0.069	-0.019	0.592	0.171
CACNG2	27	0.000	0.095	-0.087	0.000	1.000

5.3 DISCUSSION AND CONCLUSIONS

The study presented in this chapter was motivated by the question of domain-level positive selection pressure effects within tight protein complexes and the broader issue of selection pressure in functional versus non-functional regions.

I used a set of experimentally confirmed interactors of *Arc* and their binding regions' annotation as well as a wider set of proteins copurified with *Arc*. Further binding domain annotation data were sourced from Uniprot and manually filtered; also, PTM locations were acquired from Phosphosite.

I tested the difference between positive selection inside and outside binding regions of *Arc* interactors using spatial and spatiotemporal evidence from modelling. I also repeated analysis on the wider set of proteins and extended it to PTMs. *Arc* itself proved to be under overwhelmingly purifying selection pressure both in spatial and spatiotemporal analysis.

I revealed that in the case of a few of *Arc* interactors the gradual emergence of *Arc* promotes differential diversifying pressure in the regions responsible for interacting with *Arc*. This pattern is unusual in the context of a wider analysis of positive selection pressure in functionally relevant regions of proteins. Finally, results for PTMs demonstrated interesting differences in the proportions of positively selected sites between different classes of modifications.

5.3.1 *Comparison between site-specific selection methods*

As demonstrated in a brief comparison in section 5.2.3 MEME has superior recall of positively selected sites compared to FEL. The question of a possible increased false positive rate is open and would require a systematic study based on simulations. However, the general principle of a mixed effects approach to modelling (MEME is an example of such model class) has been applied in other fields too, for example, in genome-wide association studies mixed effects methods achieve superior performance while implicitly correcting for hidden structure in data (Yang et al., 2014).

5.3.2 *Specific biological effects for experimental validation*

5.3.2.1 *Arc interactors*

Based on their distinct profiles in temporal (Figure 5.5), spatial (Table 5.5), and spatio-temporal (Figure 5.8) analyses I identified 3 proteins of particular interest: *WASF1* (also knowns as *WAVE1*), *PSEN1*, and *DNM2*. Additionally, even though the spatial analysis of *Arc* was inconclusive; not only is the information about corresponding binding regions of *Arc* available for them but they also bind to three different structural units of *Arc* protein (compare binding region and structural regions in Table 5.3).

Dynamin-2 has been mentioned already in the lists of interesting genes in Chapter 3. It is a multifunctional protein interacting with multiple partners in the endocytosis process which can explain why not all peaks of episodic protein-wide diversification are related to its place in *Arc* complex. For example, its most recent diversification peak (*Homo Sapiens* branch - see Figure 5.5) is not reflected in the *Arc* binding domain (Figure 5.8a); hence, at that branch, positive sites are relatively under-represented compared to the rest of the protein sequence.

This observation is consistent with the idea of domain-level selection pressure events. Even if interaction with *Arc* is not linked to this recent episodic event in *Dynamin-2*, there might be another protein in *Arc* complex which through its appearance or active diversification in the complex induces positive selection response in *Dynamin-2*. Most likely it is not any of the already confirmed interactors, however, there are plenty of co-purified proteins (the wider set of *Arc* complex) with evidence for protein-wide episodic positive selection in recent branches (see right side of the heatmap in Figure 5.4).

On the other hand, in that context, the evidence for ongoing diversification in *Presenilin* in branches more recent than for other interactors, specifically in its region binding *Arc*, seems out of place (Figure 5.8), especially due to lack evidence for recent positive diversification (synchronised with *PSEN1* events) in that region of *Arc* (Figure 5.9). It makes it more difficult to associate the ongoing positive selection in that domain of *PSEN1* to protein's role in *Arc* complex. However, *Presenilin* binds to the N-terminus domain of *Arc*, which is less well-studied with respect to its functional, structural, and evolutionary characteristics.

However, it is equally possible that the region of *PSEN1* annotated as interacting with *Arc* plays a role in other functions, and that was the source of a more recent diversifying pressure event. Research into *Presenilin* focussed on its relationship with Alzheimer's disease and confirmed its role in the early onset type of the disease (Lee et al., 2010). In the context of the transmembrane role of this protein the region close to N terminus of *Presenilin-1* overlapping with the binding region was found to be a cytoplasmic region (Lehmann et al., 1997). This region, however, is not particularly rich in mutations associated with Alzheimer's disease - out of 95 variants listed in Uniprot aggregated from 40 publications only one overlaps with the region of interest - amino acid number 35 (Rogaeva et al., 2001). My preliminary search for binding regions (see section 5.2.4.1) identified only 3 other defined binding regions, none of them overlapping with *Arc* binding region, however, as mentioned previously, this does not exclude the possibility that *Arc* binding region is involved in other interactions too due to sparsity of annotations.

Overall, I propose that the nature of interaction between *Presenilin* and *Arc*, as well as *Presenilin's* other interactions within the same functional complex should be studied in more detail.

Finally, despite the lack of conclusive evidence for protein-wide episodic diversification of *WASF1* on any branches (see Figure 5.5) the *Arc* binding domain of *WASF1* follows a very similar diversification timeline to *Dynamin's* domain in Figure 5.8 despite targeting a slightly different region of *Arc*. *WASF1* targets the N-lobe, whereas for *DNM2* it is the end of the central linker and the beginning of the N-lobe. Functionally, in the context of *Arc* role at the synapse, *WASF1* participates in a different functional pathway to *DNM2* and *PSEN1*. The latter two take part in endocytic trafficking - AMPA receptor endocytocis and γ-secretase complex, both cascades leading to denditic spine shrinking whereas *WASF1* participates in spine F-Actin formation which contributes to LTP-related dendritic spine growth (Insall and Machesky, 2009). The regulation of spine shrinkage and spine growth mechanisms needs be balanced hence it appears that the synchronised evolution of interactions implied in both mechanisms can be justified.

5.3.2.2 *Systematic trends*

An attempt at a systematic study of the inside-domain vs. outside-domain selection in the wide set of *Arc* complex proteins revealed how unusual the positively selected domains in Arc interactors are. Most domains are under purifying pressure compared to the surrounding regions yet a select few occupy the tail of the distribution. Further to that, the link between different kinds of PTMs and positive selection presented in section 5.2.8 gives a further indication that spatial features of proteins evolve differently. Differences between the frequencies of positive sites for different classes of modifications are not easy to interpret but further integrative analysis leveraging spatial and spatio-temporal modelling paradigms employed through this thesis may suggest likely candidates for further investigation.

5.3.2.3 *Arc*

Throughout this entire thesis I have tended to focus on evidence for diversifying pressure, however, a lack of such evidence is also informative about protein. *Arc* provides a great example as its sequence was found to be under purifying pressure in spatial modelling as well as in temporal modelling paradigms. Following the interpretation of the most recent diversification postulated earlier, the extent of negative selection past early mammalian branches in *Arc* points to its optimised role within the local ecosystem of interacting proteins as well as to its finely tuned regulatory function, as any non-synonymous mutations were selected against. The only exception seems to

be the linker region which was the only place which gave any convincing indication of diversifying pressure events.

It was the case for both the unannotated part and the portion involved in *DNM2* binding as illustrated by Figure 5.9. Further to *DNM2* interaction region, the stretch of the protein encompassing the end of the linker region and the edge of N-lobe also plays a role in *Adaptor protein complex subunits* (*AP2A2*, *AP2B1*, *AP2S1*, *AP2M1*) interaction. It could not have been studied on its own as it is localised to a single amino acid site. Adaptor protein complex is involved in *Clathrin* dependent endocytosis (see section 1.2.3) and more generally in vescicle mediated transport. Hence, overall, this region of *Arc* appears to influence endocytosis through 2 interactions - with *Dynamin* as well as with various *AP2* subunits. It remains unclear why this particular function would have been associated with more positive selection than others.

Also, *Serine* at position 206 is phosphorylated by *ERK* during LTP (Nikolaienko et al., in press) (see section 1.2.3.3 for an explanation of significance of this process). Positive selection in that region in relatively recent branches fits with the discussion of refinement in complex kinase signalling cascades as demonstrated with GPCR signalling network in Chapter 4. Although the evidence is not conclusive we can theorise that in this case the fine balance of multiple cascades of phosphorylation took longer to achieve (positive selection up until early mammalian branches) compared to the remaining functions of *Arc* (purifying selection across the available depth of the evolution path).

5.3.3 *Potential for identifying more interactors*

Proteins identified as differentially copurifying with *Arc* could be its potential interactors or members of the complex through indirect relationship with *Arc*. Although the spatial and temporal effect of *Arc* interactors' binding domains is present, there are a few interactors for which I did not observe it, and on its own it cannot be considered an argument for protein's relationship with *Arc*. Findings about hubs in the previous chapter suggest influence from another, more functionally relevant interactor can dominate the effect of *Arc* appearance and initial diversification on complex members.

Therefore, here I conclude that with the available modelling data there is no reliable way of identifying further putative functional interactors. However, in a setting with better temporal resolution of divergence points and more complete interaction

data (including binding regions) I can imagine that a question such as this could be tackled by modelling signal propagation in the network of proteins over time. Possibly future work can address this issue and provide further support for the utility of evolutionary profiling of proteins in the context of discovery of novel functional interactions.

5.3.4 *Structural effects hypothesis*

Here I focussed on regions directly associated with binding other proteins, however, binding site accessibility can easily be affected by a change in a distant site of the protein which happens to affect the secondary structure.
Overall, I avoided inference based on single isolated sites (hence exclusion of short binding regions and pooling PTMs together). However, theoretically a close correlation of a binding region's temporal selection profile with a profile of a different short stretch of protein could indicate a functional link between these two regions - either regulation of the binding region, or protein self-association. The caveat of this approach is the ability to confidently assess selection pressure at single site level. Even the authors of the MEME method warn against the over-interpretation of single site evidence (Murrell et al., 2012).

5.3.5 *Implications for temporal hub effects*

In the previous chapter I proposed the idea of local hubs extending influence over their neighbouring proteins, especially at the point of their emergence. This chapter's observations allow me to revise this hypothesis, and offer a possible explanation for the noise and relatively low magnitude of effects in sections 4.2.5 and 4.3.3.

First, as discussed previously, the appearance of a protein is not necessarily a sudden event. For the purpose of a timeline discretised by divergence points such as here, protein appearance can be sudden if there were no orthologs at one divergence point, and the ortholog associated with following divergence point is already similar enough to the human protein to be picked up by the classification method. However, in many cases the situation will resemble *Arc's* appearance where there is evidence for earlier (partial) orthology yet only *X. tropicalis* is classified as the ortholog; resulting in origin at point 17 instead of 19 if we were to accept *D. rerio* ortholog or even at

22 if the partial ortholog in *D. melanogaster* was accepted.

In a large scale analysis a line needs to be drawn for which ortholog is the deepest, and this is the approach I used for my analyses in the previous chapter which allowed me to test effects at a systematic scale.
Results for *Arc* and its interactors demonstrate how much diversification in interactors happened prior to the consensus appearance of *Arc* as determined by Ensembl Compara. Limiting our analysis to branches following divergence point 17 offers a very limited and misleading picture of the true hub effect of *Arc*. A similar situation for other proteins studied as late-appearing hubs in previous chapters may add noise and reduce observable effect when investigating effect of hub appearance and diversification.

Second, as opposed to using protein-wide evidence for positive selection at a branch in chapters 3, and 4, here I looked at positive selection pressure events affecting parts of the protein, specifically analysing the difference between frequencies of positive sites in and outside the region of interest. As discussed in section 5.3.2, *Dynamin-2* provides a good example where region-specific analysis reveals how protein-wide evidence (at the *Homo Sapiens* branch) can be misleading. The branch is identified as being under episodic positive selection but it is irrelevant to the interaction of interest.
The inverse situation can also happen. Although aBSREL modelling approach does not require a large proportion of sites to belong to the positive selection category in order for the positive selection model to be favoured at a given branch. The method's authors discuss that a very small proportion can affect the method's recall of positive branches (Smith et al., 2015) so highly localised selection pressure events of modest magnitude can be completely missed by that method. These false negatives and false positives in protein-wide data can further occlude the true effect size.

As a conclusion, I propose that if adequate systematic annotation of binding regions as well as reliable information about the gradual emergence of proteins were available then the hypothesis about propagation of positive episodic selection could be extended to domain level and tested with better precision.

6

GENERAL DISCUSSION

The research project described in this thesis was set out to broaden our understanding of the molecular level of protein evolution at synapses and its relationship to observable, phenotypic changes through computational modelling while leveraging a wealth of multi-species sequence data and a variety of annotation data classes.

Following introduction of the biological background and description of the motivation for my research in Chapter 1, I reviewed the methodological principles of phylogenetic modelling and described assembly of the modelling workflow in Chapter 2. Then, Chapter 3 focussed on exploratory analysis of temporal selection profiles of the full human proteome and specifically synaptic proteins. I revealed the broad picture of emergence and main waves of episodic diversification, then identified measures for comparison of proteins but the analysis suffered from insufficient power to detect associations between function and evolutionary profile through gene set enrichment analysis. I addressed these issues in Chapter 4 where I integrated interactome data and pathway annotations to show the differences in evolutionary timeline profiles for different pathways and individual proteins based on their role in the interactome and specific pathways. I also formulated a hypothesis about the effect of appearance and diversification of proteins in tightly co-dependent molecular complexes. I further tested and revised this theory in Chapter 5 after integrating spatial as well as spatio-temporal modelling results for a small subset of proteins involved in synaptic plasticity. Finally, I proposed where my observations fit with the broader biological setting at the molecular as well as cognitive level.

Here, I bring evidence from all previous chapters together, summarise main methodological and biological contributions of the thesis, discuss sources of this research project's limitations and potential to overcome them, and finally, I propose ideas for future research in the field.

6.1 CONTRIBUTIONS SUMMARY

My contributions can be broadly divided into methodological findings including mechanistic explanations of systematic effects, and biological applications, novel information about pathways and specific proteins.

From the methodological standpoint I demonstrated the utility of complex molecular phylogenetic modelling methods in application to large sets of varied proteins across large evolutionary distances.
Based on data at the scale of the full proteome I formulated hypotheses about systematic effect linking selection pressure events and proteins' interactions, I further refined my theory thanks to a different class of data - functional pathways. Finally, I proposed a mechanistic explanation at the level of binding domains and demonstrated its utility in a small set of synaptic proteins - *Arc* complex.

By applying my modelling work specifically to the nervous system and key synaptic pathways, I improved our understanding of how molecular processes lead to phenotypic changes in the nervous system over the course of evolution, extended and revised earlier work of Emes et al. (2008) and Bayés et al. (2012). Based on comparative pathway-based analysis I postulated the major role that GPCR signalling played in refining synaptic function after establishing the basic function in simple eukaryotes. Feature extraction and clustering of modelling results allowed me to propose a selection of proteins for further study due to their extreme evolution patterns. Further to that, in the case of a few proteins in *Arc* complex I was able to discuss likely selection events for regions involved in binding.

In a broader sense, through all of the above, I contributed to the ongoing discussion about the molecular origin of human cognition.
It is a pioneering work using modelling techniques of high computational complexity at the unprecedented scale of the full proteome. Not only does it extend earlier studies in the field by providing much higher granularity of results but also it provides completely novel hypotheses and insights into molecular mechanisms of protein evolution.

6.2 LIMITATIONS OF THE STUDY

6.2.1 *Coding sequence*

Firstly, as mentioned throughout chapter 2, my methodological framework focussed exclusively on the coding sequence - transcripts and the resulting proteins. Non-coding parts of the DNA: introns, promoters and enhancers have a remarkable regulatory function. Specifically, for the synaptic function, transcription is an important part of the late LTP, thus the evolution of the regulatory role of these regions could provide complementary explanation for the phenotypic changes in the complex nervous system function (see Villar et al., 2015, for an example of an enhancer evolution study).
Further to that, recently, many classes of non-coding RNA have been described. They are transcribed from DNA and play a regulatory role in their RNA form (often attached to protein complexes). microRNA (miRNA) are perhaps the best studied of them, and their regulatory potential for LTP was already demonstrated through differential expression experiments following in-vivo LTP (Pai et al., 2014). In both cases the difficulty lies in the conceptualisation of selection pressure as there is no unambiguous way of distinguishing synonymous mutations which would allow application of the canonical definition of selection pressure (see 2.1.7) and associated methodological framework of its inference.

6.2.2 *Pathway data*

As mentioned previously, links in a pathway represent various actions such as phosphorylation, forming complexes, activation, or inhibition. Then, when reducing pathway to a graph I stripped all information about the state of the node related to these inter-node actions.
A reliable and systematic annotation of pathways in their graph form would allow for an analysis differentiating various classes of graph edges. Node states could be modelled as separate nodes, yet the mechanistic molecular relationship between such differentiation and the node diversification profile is unclear.
Also, pathways adapted through time and the fact that I only used orthologs of human proteins is limiting. There could be proteins present in other taxa pathways

which do not have human ortholog, or vice versa, a human ortholog is not a pathway member. This affects any analysis concentrating on mere binary presence or absence of proteins more than the approach based on selection pressure estimates which I adopted here. Even in my approach it can also introduce noise through inclusion of pathway members which should not be there (or exclusion of pathway members not encapsulated through orthology on its own).
Again, this issue comes down to the annotation, it is a challenge to experimentally study pathways in each sequenced organism separately searching for small differences. Perhaps computational modelling data such as mine can feed back to pathway annotation, helping to prioritise study of specific pathways in certain taxa.

6.2.3 *Interaction data*

Similar to the case of pathway data limitations discussed above, higher quality and more detailed annotation of physical protein-protein interactions would open possibilities for more insightful and precise study of the relationship to the diversification timeline. A reliable annotation of specific regions involved in interactions, as well as necessary and sufficient environmental conditions for the interaction, is lacking at the moment; and commonly used high-throughput interaction studies are unlikely to provide the data. At the same time, low-throughput studies' strength in the focus on a handful of proteins becomes their weakness when we take the scale of the full protein-protein interactome into consideration.

Moreover, as already discussed in section 5.3.4, the effects playing role in protein-protein interactions are complex and may reach beyond primary sequence proximity. In the context of spatial influence of binding proteins, it could be only a handful of amino acids within the binding region which decide its suitability for interaction and binding characteristics. Abstraction over wider regions allows for statistical inference with higher confidence but can miss biologically valid effects.

6.3 FURTHER RESEARCH IDEAS

While answering the initial research questions I realised the potential and limitations of modelling techniques, I was able to formulate new questions for future investigation. In this section I briefly draw ideas for further work in the field of large-scale

phylogenetic inference as well as more specifically studying conservation and diversification of PSD.

6.3.1 *Phylogenetic methodology research*

First, I postulate the use of simulations for further exploration of the effect of different methods of spatial selection pressure inference (section 5.2.3), as well as the effect of limiting the number of taxa (section 5.2.9).

A common procedure in molecular phylogenetic work is to use all available ortholog sequences of a target gene to construct a MSA. The number of annotated genomes grows constantly and this trend is likely to accelerate. However, distribution of available annotated genomes on the tree of life is biased towards complex vertebrate organisms, especially mammals, which in turn biases the availability of orthologs for MSA.
Here, the goal would be to quantify the influence of this bias on the downstream phylogenetic analysis, specifically on site specific selection pressure estimates. In parallel the effect of eliminating poorly aligned columns of MSA could be tested, which had been previously shown to improve accuracy of the analysis (Talavera and Castresana, 2007; Chang et al., 2014).
The overarching question is whether reducing the size of MSA in both dimensions can improve its informativeness for estimating selection pressure enough to outweigh the loss of power of the test.

In order to have access to gold standard data about the parameters of sequence evolution we are forced to use sequence evolution simulators. Their biological relevance is debatable because they are limited by our simplified mathematical models of molecular evolution, and blind to functional significance of the simulated sequence.
The minimum requirement for sequence simulation algorithm is to account for mutation rates, insertions and deletions. For specific use cases such as this project we can expect the simulator to account for open reading frame, selection pressure and allow the user to vary parameters along the sequence to simulate the distinction between functional domains and non-functional inter-domain regions.
There are multiple frameworks available, however, not all of them satisfy all criteria which I had identified.

For example a few older programs such as `indel-Seq-Gen` (Strope et al., 2007) and `Dawg` (Cartwright, 2005) do not support codon evolution models. Amongst newer frameworks, such as `INDELible` (Fletcher and Yang, 2009), `EvolveAGene` (Hall, 2008), or `PhyloSim` (Sipos et al., 2011), the last one is the best choice at the moment. It is implemented as `R` package, thus integrates well with the supporting analysis code, and allows for an easy extension of the basic use case.

Validity of simulated data can be improved by simulation parameters fitting with real-world data so the output sequences at the nodes of the tree follow the same distribution of lengths and sequence similarity as naturally evolved sequences.

6.3.2 *Summary measures*

Throughout the thesis I used different ways of summarising the evolution of a protein - discretising and binning temporal distributions - extracting summary measures, using raw and smoothed vectors for distance metric.
Then in Chapter 5 I counted positive sites across stretches of a protein, and discussed it as a proxy for how much positive selection pressure was experienced by a region. However, as mentioned in section 5.3.4, this is just an approximation of the real impact of evolution on domain's functionality.

The outstanding issue is whether the effect of diversifying and purifying pressure on a protein can be summarised in one number and how spatial summaries relate to temporal measures. Here, in Figure 6.1, I contrast two very crude measures across all proteins - the relative number of positive branches and the relative number of positive sites - although the data are noisy, they seem to correlate weakly. Further methodological research could also focus on the accurate way of capturing region-specific selection magnitude from full protein results, which does not require separate region-specific alignment and full modelling workflow execution.

6.3.3 *Systematic study of binding domains co-diversification*

The ideas discussed in the previous section link directly with an extension of my pilot inquiry into the propagation of positive selection events between binding regions.
When I used *Arc* complex I discovered two main obstacles:

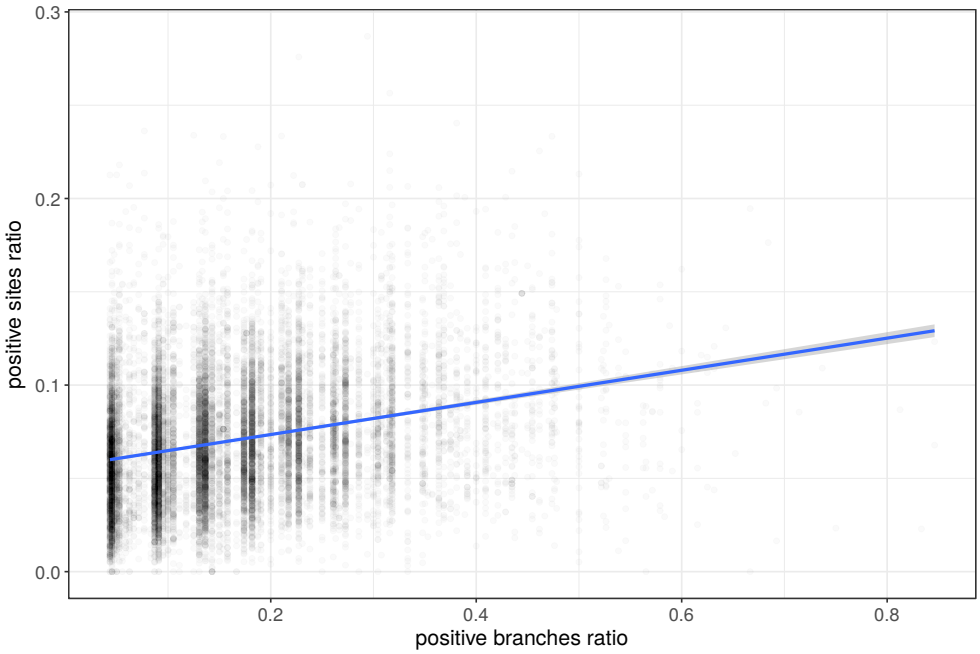


Figure 6.1: Relationship between site-based summary and branch-based summary of selection pressure.

1. Practically no spatio-temporal pattern of diversification in *Arc* itself - generally the entire sequence is under purifying pressure.

2. Only a single one-to-many group investigated, interference from other proteins which bind to *Arc* interactors at regions overlapping with *Arc*-binding ones was intractable. However, I discussed it as a possible explanation of temporal effects observed for *DNM2* and *PSEN1* and unaccounted for by *Arc* interaction on its own (see section 5.3.2.1).

An extension of the same approach would require high confidence data about interaction sites. At the moment, to the best of my knowledge, it would also require manual curation of data which is prohibitively expensive even at a scale of PSP.
This line of inquiry can be extended further if we integrate domain-specific episodic selection patterns across a large group of interacting proteins and train a classifier to reconstruct the network topology in a similar fashion as gene interaction networks can be reconstructed from temporal expression data (Gardner, 2003).
Should the results of this test be successful, i.e. network topology could be recognised substantially better than chance level, temporal evolutionary data could be used for putative interactor identification, as well as for narrowing down interactor lists to the functional ones.

6.3.4 *Focus on other species*

Here, in any analyses with a temporal component, my attention was focussed on the path leading to *h. sapiens* leaf. Equally, analysis could easily focus on the path to another node, the primary examples would be model organisms used in neuroscience research, such as mouse or rat. Both systematic effects as well as specific applications could be studied this way. First, hub effects postulated in the thesis could be validated on different paths. Also, provided sufficient density of divergence points on other clades, temporal distribution of MRP peaks could be tested. Peaks synchronised with those on the human path would suggest external environmental factors affected all clades equally; conversely, different timing would indicate selection events had different origin, possibly unrelated to Earth's geological history.

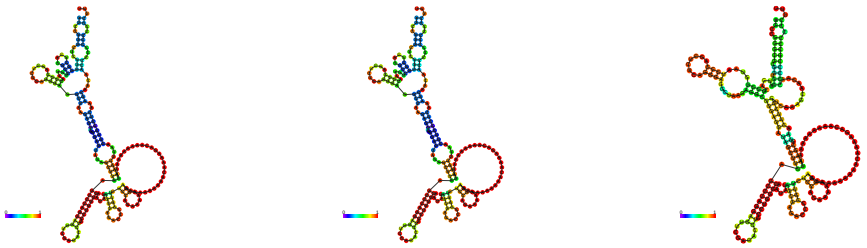
Second, the proposed inquiry presents potential for comparative analysis of how certain protein complexes evolved past the point of divergence between rodents and primates. If a functional pathway was under positive selection in primate clade but not in rodents then the validity of clinical research affecting this pathway is put into question.
Of course pathways which are more important for rodents than for primates, such as olfaction (discussed in section 4.4.2), would receive more attention, leading to improvement in our understanding of the origin of their function.
The major impact of this extension of my research would be in drug discovery process and any research into the molecular aetiology of complex disease which is carried out in model organisms.

6.3.5 *Non-coding sequence*

As explained in section 6.2.1 non-coding sequence plays substantial regulatory role, and it evolves at much higher rates that coding sequence (Rands et al., 2014). Thus, information about evolutionary profiles of non-coding sequence could contribute a wealth of novel insights into understanding of complex function origin, e.g. in the synapse.
Here, I propose a study focussed on identification of positive selection events in non-coding RNA (ncRNA) such as long non-coding RNA (lncRNA) and miRNA. The core concept common to studies of selection pressure in coding sequence is using non-

synonymous to synonymous ratio as a proxy for selection pressure measurement.
In non-coding sequence other methods, which do not rely on this conceptualisation, can be used. For example McDonald-Kreitman test allows detection of positive selection based on a comparison of inter-species variability to intra-species polymorphisms (McDonald and Kreitman, 1991), in its standard form it was conceived for coding sequence but in its generalised form it does nor require the region in question to be coding (Egea et al., 2008).
Another approach, Siphy-ω infers patterns of substitutions characteristic for selection constraints with a HMM which allows for reasoning about selection pressure without the need to use the rate of substitutions as a defining criterion (Garber et al., 2009). This method was successfully applied to a whole-genome study and allowed discovery of novel regions of the genome under selective constraints (Lindblad-Toh et al., 2011). However, if we were to transfer the coding sequence methodology to non-coding sequence, mutations can be defined as mimetic or non-mimetic with respect to any arbitrary functional characteristic in order to avoid the clash of nomenclature with the commonly accepted definition of synonymous and non-synonymous coding sequence mutations. Within this generalised framework synonymous and nonsynonymous mutations of codons are a special case where we defined the functional characteristic in terms of the aminoacid translated from a codon.
In fact, there is evidence for selection pressure in coding-sequence not based on aminoacid meaning of the codons but on their secondary structure or regulatory role - called non-coding selection pressure (Chen and Blanchette, 2007).
Moreover, studies adopting the generalised concept of a ratio of two classes of substitutions to non-coding sequence already exist. For example, in case of *cis*-regulatory regions mutations can be divided into upregulating, downregulating, or silent, so here mutations are defined as mimetic with respect to the expression level of the gene regulated by the region (Smith et al., 2013). This definition, together with data about effects of arbitrary mutations from high-throughput mutagenesis studies (e.g. Melnikov et al., 2012), allowed the authors to demonstrate utility of the framework on a selection of enhancers: *LTV1*, *ALDOB*, and *ECR11*.

The secondary structure of lncRNA was found to be functionally relevant (Novikova et al., 2012) and discussed as evolutionary conserved (Torarinsson et al., 2006) , even though the lncRNA sequence is often described as poorly conserved (Johnsson et al., 2014), Diederichs (2014) takes this argument a step further and claims that full description of lncRNA evolution can only be achieved with a combination of sequence,



(a) original sequence (b) synonymous mutation (c) non-synonymous mutation

Figure 6.2: *BC200* lncRNA minimum free energy secondary structure predicted with `RNAfold`, (a) orginal sequence, (b) single synonymous mutation introduced - same structure, (c) single non-synonymous mutation introduced - different structure

structure, function, and expression features. However, one could make this argument for any sequence, coding or non-coding. In practice we need to compromise certain biological factors for simplicity of the model and choose only the most informative features.

Emerging approaches use secondary structure predictions to guide selection pressure analysis of lncRNA (M. Costa & K. Nowick, personal communication). The authors used `RNAfold`, part of `ViennaRNA` software package (Lorenz et al., 2011), to predict the most likely secondary structure (see Figure 6.2a) and `RNAsnp` to detect changes in folding induced by mutations (Sabarinathan et al., 2013). Selection pressure is estimated based on the assumption that mimetic mutations result in the same secondary structure (see Figure 6.2b) while non-mimetic ones change it (see Figure 6.2c).

Alternative methods which could be integrated in a similar framework include `RNAstructure` package (Mathews, 2006) and `MultiAlign` which estimates a consensus structure for multiple homologous RNA sequences (Xu and Mathews, 2011), lncRNA sequences can be sourced from lncpedia (Volders et al., 2014).

Also, the same principle can be extended to miRNA or its intermediate forms such as pri-miRNA or pre-miRNA as miRNA has a widespread regulatory role which is the subject of ongoing research (Lin and Gregory, 2015; Ameres and Zamore, 2013).

If the assembly of a methodological framework is successful and modelling results are calculated for ncRNA known to be involved in synaptic regulation, I would expect to find evidence for late or ongoing diversification of ncRNA which could point to their contribution to the emergence of complex synaptic function. It could also facilitate the identification of candidates for experimental functional validation among sequences which are less thoroughly characterised.

6.4 FINAL REMARKS

On top of the specific immediate scientific contributions described earlier in section 6.1, the work reported in this thesis has potential for further impact beyond this research project.

Firstly, the long-lasting legacy of this project lies in a flexible modelling and analytic workflow which can be reapplied to different taxonomic selections of orthologs, and to completely different paradigms of sequence data acquisition, for instance, directly from experiments in bacteria. Also, I am leaving behind a rich modelling output data, a product of many compute-hours, which can be further analysed as better sequence annotation data become available (e.g. as discussed in section 6.3.3).
Finally, I hope researchers will find global systematic hypotheses and specific leads for further experimental work postulated in this thesis inspiring for further research in this field.

BIBLIOGRAPHY

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. pages 420–434. (Cited on page 55.)

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*. (Cited on page 24.)

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607. (Cited on pages 46, 47, and 76.)

Allen, B. L. and Steel, M. (2001). Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*, 5(1):1–15. (Cited on pages 26 and 41.)

Altschul, S. F., Madden, T. L., Schäffer, a. a., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402. (Cited on page 19.)

Ameres, S. L. and Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nature reviews. Molecular cell biology*, 14(8):475–88. (Cited on page 170.)

Anagnostaras, S. G., Murphy, G. G., Hamilton, S. E., Mitchell, S. L., Rahnama, N. P., Nathanson, N. M., and Silva, A. J. (2003). Selective cognitive dysfunction in acetylcholine M1 muscarinic receptor mutant mice. *Nature neuroscience*, 6(1):51–58. (Cited on page 11.)

Ariew, R. (1976). *Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony*. University of Illinois. (Cited on page 26.)

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29. (Cited on page 45.)

Baier, H. and Bonhoeffer, F. (1992). Axon Guidance by Gradients of a Target-Derived Component. *Science*, 255:472–475. (Cited on page 5.)

Banko, J. L. (2006). Regulation of Eukaryotic Initiation Factor 4E by Converging Signaling Pathways during Metabotropic Glutamate Receptor-Dependent Long-Term Depression. *Journal of Neuroscience*, 26(8):2167–2173. (Cited on page 10.)

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., Macdougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sun-

daram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., Yerramalla, M. S., and Zhang, J. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212. (Cited on pages 34 and 140.)

Bayés, A., Collins, M. O., Croning, M. D. R., van de Lagemaat, L. N., Choudhary, J. S., and Grant, S. G. N. (2012). Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. *PloS one*, 7(10):e46683. (Cited on pages 12, 44, and 162.)

Bayés, A., van de Lagemaat, L. N., Collins, M. O., Croning, M. D. R., Whittle, I. R., Choudhary, J. S., and Grant, S. G. N. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1):19–21. (Cited on page 3.)

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ. (Cited on page 22.)

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 Published by :. *J R Statist Soc B*, 57(1):289–300. (Cited on page 47.)

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Burgess, S., Buza, T., Gresham, C., McCarthy, F., Pillai, L., Wang, H., Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Feuermann, M., Gaudet, P., Basu, S., Chisholm, R. L., Dodson, R. J., Fey, P., Mi, H., Thomas, P. D., Muruganujan, A., Poudel, S., Hu, J. C., Aleksander, S. A., McIntosh, B. K., Renfro, D. P., Siegele, D. A., Attrill, H., Brown, N. H., Tweedie, S., Lomax, J., Osumi-Sutherland, D., Parkinson, H., Roncaglia, P., Lovering, R. C., Talmud, P. J., Humphries, S. E., Denny, P., Campbell, N. H., Foulger, R. E., Chibucos, M. C., Giglio, M. G., Chang, H. Y., Finn, R., Fraser, M., Mitchell, A., Nuka, G., Pesseat, S., Sangrador, A., Scheremetjew, M., Young, S. Y., Stephan, R., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Bahler, J., Lock, A., Kersey, P. J., McDowall, M. D., Staines, D. M., Dwinell, M., Shimoyama, M., Laulederkind, S., Hayman, G. T., Wang, S. J., Petri, V., D'Eustachio, P., Matthews, L., Balakrishnan, R., Binkley, G., Cherry, J. M., Costanzo, M. C., Demeter, J., Dwight, S. S., Engel, S. R., Hitz, B. C., Inglis, D. O., Lloyd, P., Miyasato, S. R., Paskov, K., Roe, G., Simison, M., Nash, R. S., Skrzypek, M. S., Weng, S., Wong, E. D., Berardini, T. Z., Li, D., Huala, E., Argasinska, J., Arighi, C., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Bely, B., Blatter, M. C., Bonilla, C., Bougueleret, L., Boutet, E., Breuza, L., Bridge, A., Britto, R., Casals, C., Cibrian-Uhalte, E., Coudert, E., Cusin, I., Duek-Roggli, P., Estreicher, A., Famiglietti, L., Gane, P., Garmiri, P., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Hulo, C., Huntley, R., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Macdougall, A., Magrane, M., Martin, M., Masson, P., Mutowo, P., O'Donovan, C., Pedruzzi, I., Pichler, K., Poggioli, D., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Wu, C., Xenarios, I., Chan, J., Kishore, R., Sternberg, P. W., Van Auken, K., Muller, H. M., Done, J., Li, Y., Howe, D., and Westerfeld, M. (2015). Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056. (Cited on page 45.)

Bradshaw, J. L. and Rogers, L. J. (1993). *The evolution of lateral asymmetries, language, tool use, and intellect.* Academic Press. (Cited on page 81.)

Bramham, C. R., Alme, M. N., Bittins, M., Kuipers, S. D., Nair, R. R., Pai, B., Panja, D., Schubert, M., Soule, J., Tiron, A., and Wibrand, K. (2010). The Arc of synaptic memory. *Experimental brain research*, 200(2):125–40. (Cited on page 131.)

Bramham, C. R., Worley, P. F., Moore, M. J., and Guzowski, J. F. (2008). The immediate early gene Arc/Arg3.1. *Journal of Neuroscience*, 28(46):11760–11767. (Cited on page 131.)

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192. (Cited on page 118.)

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10:1–9. (Cited on page 20.)

Campillos, M., Doerks, T., Shah, P. K., and Bork, P. (2006). Computational characterization of multiple Gag-like human proteins. *Trends in Genetics*, 22(11):585–589. (Cited on pages 132 and 133.)

Cao, C., Rioult-pedotti, M. S., Migani, P., Yu, C. J., Tiwari, R., Parang, K., Spaller, M. R., Goebel, D. J., and Marshall, J. (2013). Impairment of TrkB-PSD-95 Signaling in Angelman Syndrome. *PLoS biology*, 11(2). (Cited on page 135.)

Caporale, N. and Dan, Y. (2008). Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience*, 31(1):25–46. (Cited on page 8.)

Cartwright, R. a. (2005). DNA assembly with gaps (Dawg): Simulating sequence evolution. *Bioinformatics*, 21(SUPPL. 3):31–38. (Cited on page 166.)

Castellucci, V. and Kandel, E. (1976). Presynaptic facilitation as a mechanism for behavioral sensitization in Aplysia. *Science*, 194(4270):1176–1178. (Cited on page 11.)

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 12(8):773–777. (Cited on page 45.)

Chang, J. M., Di Tommaso, P., and Notredame, C. (2014). TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 31(6):1625–1637. (Cited on pages 34 and 165.)

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B. J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379. (Cited on page 91.)

Chen, H. and Blanchette, M. (2007). Detecting non-coding selective pressure in coding regions. *BMC evolutionary biology*, 7 Suppl 1:S9. (Cited on page 169.)

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335. (Cited on page 27.)

Collingride, G. and Bliss, T. V. P. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*. (Cited on page 10.)

Collins, M. O., Husi, H., Yu, L., Brandon, J. M., Anderson, C. N. G., Blackstock, W. P., Choudhary, J. S., and Grant, S. G. N. (2006). Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of neurochemistry*, 97 Suppl 1(m):16–23. (Cited on page 44.)

Costa-Mattioli, M., Gobert, D., Harding, H., Herdy, B., Azzi, M., Bruno, M., Bidinosti, M., Ben Mamou, C., Marcinkiewicz, E., Yoshida, M., Imataka, H., Cuello, A. C., Seidah, N., Sossin, W., Lacaille, J.-C., Ron, D., Nader, K., and Sonenberg, N. (2005). Translational control of hippocampal synaptic plasticity and memory by the eIF2alpha kinase GCN2. *Nature*, 436(7054):1166–73. (Cited on pages 10 and 11.)

Cox, D. R. and Isham, V. (1980). *Point processes*, volume 12. CRC Press. (Cited on page 23.)

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, a. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M. J., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669. (Cited on pages 20, 33, and 34.)

Daraselia, N., Yuryev, A., Egorov, S., Mazo, I., and Ispolatov, I. (2007). Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. *BMC bioinformatics*, 8:243. (Cited on page 87.)

Dayhoff, M. O. and Schwartz, R. M. (1978). A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*. Citeseer. (Cited on page 21.)

De Camilli, P. and Jahn, R. (1990). Pathways to Regulated Exocytosis in Neurons. *Annual Review of Physiology*, 52(1):625–645. (Cited on page 81.)

Delghandi, M. P., Johannessen, M., and Moens, U. (2005). The cAMP signalling pathway activates CREB through PKA, p38 and MSK1 in NIH 3T3 cells. *Cellular Signalling*, 17(11):1343–1351. (Cited on pages 8 and 9.)

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, 30(4):121–123. (Cited on page 169.)

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2):330–340. (Cited on page 22.)

Domes, G., Heinrichs, M., Michel, A., Berger, C., and Herpertz, S. C. (2007). Oxytocin Improves "Mind-Reading" in Humans. *Biological Psychiatry*, 61(6):731–733. (Cited on page 118.)

Dorsam, R. T. and Gutkind, J. S. (2007). G-protein-coupled receptors and cancer. *Nature Reviews Cancer*, 7(2):79–94. (Cited on page 8.)

Dosemeci, A., Weinberg, R. J., Reese, T. S., and Tao-Cheng, J.-H. (2016). The Postsynaptic Density: There Is More than Meets the Eye. *Frontiers in Synaptic Neuroscience*, 8(August):1–8. (Cited on page 4.)

Dray, S. and Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4). (Cited on page 57.)

Eaton, E. and Mansbach, R. (2012). A Spin-Glass Model for Semi-Supervised Community Detection. *AAAI*, pages 900–906. (Cited on page 87.)

Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Current opinion in structural biology*, 16(3):368–73. (Cited on page 20.)

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., Spritz, R. A., Deriel, J. K., Forget, B. G., Weissman, M., Slightom, J. L., Blechl, A. E., Baralle, E., Shoulders, C. C., and Proudfootq, N. J. (1980). The Structure and Evolution of the Human Beta-Globin Gene Family. *Cell*, 21(October):653–666. (Cited on page 41.)

Egea, R., Casillas, S., and Barbadilla, A. (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic acids research*, 36(Web Server issue):157–162. (Cited on page 169.)

Emes, R., Pocklington, A., Andreson, C., Bayes, A., Collins, M., Vickers, C., Croning, M., Malik, B., Choudhary, J. S., Armstrong, J. D., and Grant, S. G. N. (2008). Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nature . . .*, 11(7):799–806. (Cited on pages 11, 13, 90, 102, and 162.)

Emes, R. D. and Grant, S. G. N. (2012). Evolution of synapse complexity and diversity. *Annual review of neuroscience*, 35:111–31. (Cited on pages 11 and 90.)

Engle, E. C. (2010). Human Genetic Disorders of Axon Guidance. *Cold Spring Harbour Perspectives in Biology*, pages 1–19. (Cited on page 5.)

Esteban, J. a., Shi, S.-H., Wilson, C., Nuriya, M., Huganir, R. L., and Malinow, R. (2003). PKA phosphorylation of AMPA receptor subunits controls synaptic trafficking underlying plasticity. *Nature neuroscience*, 6(2):136–143. (Cited on pages 9 and 117.)

Farinas, I., Jones, K. R., Backus, C., Wang, X.-Y., and Reichardt, L. F. (1994). Severe sensory and sympathetic deficits in mice lacking neurotrophin-3. *Nature*, 369(6482):658–661. (Cited on page 5.)

Fariñas, I., Wilkinson, G. A., Backus, C., and Reichardt, L. F. (1998). Characterization of Neurotrophin and Trk Receptor Functions in Developing Sensory Ganglia: Direct NT-3 Activation of TrkB Neurons In Vivo. *Neuron*, 21(2):325–334. (Cited on page 5.)

Federhen, S. (2003). The Taxonomy Project. In McEntyre, J. and Ostell, J., editors, *The NCBI Handbook*. NCBI, Bethesda (MD). (Cited on pages 20 and 35.)

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of molecular evolution*, 25(4):351–360. (Cited on page 20.)

Ferrara, E. (2012). A Large-Scale Community Structure Analysis In Facebook. *EPJ Data Science*, pages 1–30. (Cited on page 87.)

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(SUPPL. 2):1–9. (Cited on page 20.)

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, 19(2):99. (Cited on page 18.)

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888. (Cited on page 166.)

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41. (Cited on page 87.)

Frost, H. and McCray, A. T. (2012). Markov Chain Ontology Analysis (MCOA). *BMC Bioinformatics*, 13(1):23. (Cited on page 82.)

Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R. H., Daniel, D. J., Graham, R. L., and Woodall, T. S. (2004). Open {MPI}: Goals, Concept, and Design of a Next Generation {MPI} Implementation. In *Proceedings, 11th European PVM/MPI Users' Group Meeting*, pages 97–104, Budapest, Hungary. (Cited on page 36.)

Gallagher, M. and Hollandt, P. C. (1994). Review The amygdala complex : Multiple roles in associative learning and attention. *Proceedings of the National Academy of Sciences*, 91(December):11771–11776. (Cited on page 10.)

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):54–62. (Cited on page 169.)

Gardner, T. S. (2003). Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science*, 301(5629):102–105. (Cited on page 167.)

Gazzaniga, M. S. (2008). *Human: The science behind what makes us unique*. Ecco, New York. (Cited on page 1.)

Geifman, N., Monsonego, A., and Rubin, E. (2010). The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems. *BMC bioinformatics*, 11:458. (Cited on pages 45 and 76.)

Geyer, C. (1992). Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483. (Cited on page 27.)

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In M, K. E. and Kaufman, S. M., editors, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation of North America. (Cited on page 27.)

Gilad, Y., Man, O., and Glusman, G. (2005). A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Research*. (Cited on page 118.)

Gilad, Y., Man, O., Pä Ä Bo, S., Lancet, D., and Harpending, H. C. (2003). Human specific loss of olfactory receptor genes. *Proceedings of the National Academy of Sciences*, 100(6). (Cited on page 118.)

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, 23(8):980–987. (Cited on page 82.)

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, 11(5):725–36. (Cited on page 28.)

Goldstein, L. S. B. and Yang, Z. (2000). Microtubule-Based Transport Systems in Neurons: The Roles of Kinesins and Dyneins. *Annual Review of Neuroscience*, 23:39–71. (Cited on page 7.)

Good, I. J. and Mittal, Y. (1987). The Amalgamation and Geometry of Two-by-Two Contingency Tables. *Annals of Statistics*, 15(2):694–711. (Cited on page 120.)

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine*, 130(12):1005–13. (Cited on page 25.)

Grant, S. G. N. (2003). Synapse signalling complexes and networks: Machines underlying cognition. *BioEssays*, 25(12):1229–1235. (Cited on page 10.)

Gyles, C. and Boerlin, P. (2014). Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary pathology*, 51(2):328–40. (Cited on page 41.)

Hakes, L., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, 104(19):7999–8004. (Cited on pages 89 and 122.)

Hall, B. G. (2008). Simulating DNA coding sequence evolution with EvolveAGene 3. *Molecular Biology and Evolution*, 25(4):688–695. (Cited on page 166.)

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series*, volume 41, pages 95–98. (Cited on page 31.)

Hänggi, J., Fövenyi, L., Liem, F., Meyer, M., and Jäncke, L. (2014). The hypothesis of neuronal interconnectivity as a function of brain size-a general organization principle of the human connectome. *Frontiers in human neuroscience*, 8(November):915. (Cited on page 12.)

Harrison, P. M., Gerstein, M., and Avenue, W. (2002). Studying Genomes Through the Aeons: Protein Families, Pseudogenes and Proteome Evolution. *Journal of Molecular Biology*, 2836(02):1155–1174. (Cited on page 2.)

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97. (Cited on pages 25 and 27.)

He, X. and Simpson, T. I. (2017a). OntoSuite-Miner; an NLP pipeline for annotating ontologies from biomedical corpora. (Cited on page 76.)

He, X. and Simpson, T. I. (2017b). topOnto - An R package for generically enabled ontology enrichment analysis. (Cited on page 76.)

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919. (Cited on page 21.)

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21(1):101–114. (Cited on page 118.)

Hofman, A. (2008). Neural Networks and Cognition An Evolutionary Approach. *Japanese Journal of Cognitive Neuroscience*, 10(3):235–238. (Cited on page 12.)

Hofman, M. A. (2001). Brain evolution in hominids: are we at the end of the road. *Evolutionary anatomy of the primate cerebral cortex*, pages 113–127. (Cited on page 12.)

Hofman, M. A. (2014). Evolution of the human brain: when bigger is better. *Frontiers in neuroanatomy*, 8(March):15. (Cited on page 12.)

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., and Pääbo, S. (2001). Ancient DNA. *Nature Reviews Genetics*, 2(May):3–9. (Cited on page 1.)

Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561. (Cited on page 141.)

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*,

43(D1):D512–D520. (Cited on page 141.)

Houtman, J. C., Barda-Saad, M., and Samelson, L. E. (2005). Examining multiprotein signaling complexes from all angles: The use of complementary techniques to characterize complex formation at the adapter protein, linker for activation of T cells. *FEBS Journal*, 272(21):5426–5435. (Cited on page 89.)

Huang, C., Ni, Y., Wang, T., Gao, Y., Haudenschild, C. C., and Zhan, X. (1997). Down-regulation of the filamentous actin cross-linking activity of cortactin by Src-mediated tyrosine phosphorylation. *Journal of Biological Chemistry*, 272(21):13911–13915. (Cited on page 81.)

Hunter, T. (1995). Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell*, 80(2):225–236. (Cited on page 89.)

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., Obar, R. A., Guruharsha, K. G., Li, K., Artavanis-Tsakonas, S., Gygi, S. P., and Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509. (Cited on page 86.)

Insall, R. H. and Machesky, L. M. (2009). Actin Dynamics at the Leading Edge: From Simple Machinery to Complex Networks. *Developmental Cell*, 17(3):310–322. (Cited on page 157.)

Ispolatov, I., Mazo, I., and Yuryev, A. (2006). Finding mesoscopic communities in sparse networks. *Journal of Statistical Mechanics: ...*, pages 1–7. (Cited on page 87.)

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666. (Cited on pages 44 and 60.)

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc. (Cited on page 44.)

Janeway, C. A. and Medzhitov, R. (2002). Innate Immune Recognition. *Annual Review of Immunology*, 20(1):197–216. (Cited on page 103.)

Jensen, R. a. (2001). Orthologs and paralogs - we need to get it right. *Genome biology*, 2(8):interactions1002.1–interactions1002.3. (Cited on page 18.)

Jeong, H., Mason, S. P., Barabási, a. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–2. (Cited on page 87.)

Job, C. and Eberwine, J. (2001). Localization and translation of mRNA in dendrites and axons. *Nature reviews. Neuroscience*, 2(12):889–98. (Cited on page 9.)

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254. (Cited on pages 44 and 57.)

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta*, 1840(3):1063–71. (Cited on page 169.)

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20. (Cited on pages 85, 86, and 88.)

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33(DATABASE ISS.):428–432. (Cited on pages 45, 76, 88, and 102.)

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian protein metabolism*, pages 21–132. Academic Press, New York. (Cited on page 24.)

Kandel, E. R. (2012). The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Molecular Brain*, 5(1):14. (Cited on page 11.)

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361. (Cited on pages 46, 76, and 88.)

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30. (Cited on pages 46, 76, and 88.)

Kass, R. E., Raftery, A. E., Association, S., and Jun, N. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795. (Cited on page 146.)

Katoh, K. and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics (Oxford, England)*, 26(15):1899–900. (Cited on page 34.)

Kauer, J. a., Malenka, R. C., and Nicoll, R. a. (1988). NMDA application potentiates synaptic transmission in the hippocampus. *Nature*, 334(6179):250–252. (Cited on page 9.)

Kelleher, R. J., Govindarajan, A., and Tonegawa, S. (2004). Translational regulatory mechanisms in persistent forms of synaptic plasticity. *Neuron*, 44(1):59–73. (Cited on page 9.)

Kennedy, M. B. (2000). Signal-Processing Machines at the Postsynaptic Density. *Science*, 290(5492):750–754. (Cited on pages 3 and 7.)

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664. (Cited on page 19.)

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–97. (Cited on page 29.)

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic acids research*, 40(Database issue):D841–6. (Cited on page 91.)

Kimple, M. E., Brill, A. L., and Pasker, R. L. (2001). Overview of Affinity Tags for Protein Purification. In *Current Protocols in Protein Science*. John Wiley & Sons, Inc. (Cited on page 137.)

Kimura, M. (1984). *The neutral theory of molecular evolution*. Cambridge University Press. (Cited on page 16.)

Klann, E., Antion, M. D., Banko, J. L., and Hou, L. (2004). Synaptic plasticity and translation initiation. *Learn.Mem.*, 11(713):365–372. (Cited on page 10.)

Knudson, C. (2016). glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation. (Cited on page 151.)

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, 39(1):309–338. (Cited on pages 18 and 19.)

Kosakovsky Pond, S., Delport, W., Muse, S. V., and Scheffler, K. (2010). Correcting the bias of empirical frequency parameter estimators in codon models. *PloS one*, 5(7):e11230. (Cited on pages 28, 30, and 36.)

Kosakovsky Pond, S. L. and Frost, S. D. W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–22. (Cited on pages 29 and 36.)

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution*, 23(10):1891–901. (Cited on page 42.)

Kosik, K. S. (2009). Exploring the early origins of the synapse by comparative genomics. *Biology letters*, 5(1):108–11. (Cited on page 11.)

Kosiol, C., Vina??, T., Da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4(8). (Cited on page 118.)

Krishnan, A. and Schioth, H. B. (2015). The role of G protein-coupled receptors in the early evolution of neurotransmission and the nervous system. *Journal of Experimental Biology*, 218(4):562–571. (Cited on page 12.)

Kurland, C. G., Collins, L. J., and Penny, D. (2006). Genomics and the Irreducible Nature of Eukaryote Cells. *Science*, 312:1011–1015. (Cited on page 2.)

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–74. (Cited on page 29.)

Lee, J. H., Yu, W. H., Kumar, A., Lee, S., Mohan, P. S., Peterhoff, C. M., Wolfe, D. M., Martinez-Vicente, M., Massey, A. C., Sovak, G., Uchiyama, Y., Westaway, D., Cuervo, A. M., and Nixon, R. A. (2010). Lysosomal proteolysis and autophagy require presenilin 1 and are disrupted by Alzheimer-related PS1 mutations. *Cell*, 141(7):1146–1158. (Cited on page 156.)

Lehmann, S., Chiesa, R., and Harris, D. A. (1997). Evidence for a six-transmembrane domain structure of presenilin 1. *Journal of Biological Chemistry*, 272(18):12047–12051. (Cited on page 156.)

Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. a. (2012). A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics (Oxford, England)*, 28(24):3248–56. (Cited on page 29.)

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. *Genome research*, pages 1048–1059. (Cited on page 2.)

Li, F.-W., Villarreal, J. C., Kelly, S., Rothfels, C. J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E. M., Der, J. P., Pittermann, J., Burge, D. O., Pokorny, L., Larsson, A., Chen, T., Weststrand, S., Thomas, P., Carpenter, E., Zhang, Y., Tian, Z., Chen, L., Yan, Z., Zhu, Y., Sun, X., Wang, J., Stevenson, D. W., Crandall-Stotler, B. J., Shaw, A. J., Deyholos, M. K., Soltis, D. E., Graham, S. W., Windham, M. D., Langdale, J. A., Wong, G. K.-S., Mathews, S., and Pryer, K. M. (2014). Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences*. (Cited on page 41.)

Li, W. (1997). *Molecular evolution*. Sinauer Associates Incorporated. (Cited on page 16.)

Lin, S. and Gregory, R. I. (2015). MicroRNA biogenesis pathways in cancer. *Nature Review Cancer*, 15(6):321–333. (Cited on page 170.)

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., Kellis, M., Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Cree, A., Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., and Sodergren, E. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482. (Cited on page 169.)

Liò, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research*, 8(12):1233–1244. (Cited on page 24.)

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26. (Cited on page 170.)

Magee, J. C. and Johnston, D. (1997). A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science (New York, N.Y.)*, 275(January):209–213. (Cited on page 8.)

Mangiamele, L. a., Thomson, C. J., Lebonville, C. L., and Burmeister, S. S. (2010). Characterization of the plasticity-related gene, Arc, in the frog brain. *Developmental neurobiology*, 70(12):813–25. (Cited on page 133.)

Martin, S. J., Grimwood, P. D., and Morris, R. G. M. (2000). Synaptic plasticity and memory: An Evaluation of the Hypothesis. *Annual review of neuroscience*, (Hebb 1949):649–711. (Cited on page 10.)

Mathews, D. H. (2006). RNA secondary structure analysis using RNAstructure. *Current Protocols in Bioinformatics*, pages 12–16. (Cited on page 170.)

Mattaliano, M. D., Montana, E. S., Parisky, K. M., Littleton, J. T., and Griffith, L. C. (2007). The Drosophila ARC homolog regulates behavioral responses to starvation. *Molecular and cellular neurosciences*, 36(2):211–21. (Cited on page 132.)

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. (2007). Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics (Oxford, England)*, 23(13):i319–27. (Cited on page 29.)

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistic Association*, 92(437):162–170. (Cited on page 150.)

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654. (Cited on page 169.)

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan Jr, C. G., and Kinney, J. B. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology*, 30(3):271–277. (Cited on page 169.)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087. (Cited on page 27.)

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189. (Cited on pages 45, 76, and 88.)

Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. a. (2007). Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes. *Genetics*, 175(4):1773–85. (Cited on page 42.)

Morris, R. L. and Hollenbeck, P. J. (1995). Axonal transport of mitochondria along microtubules and F-actin in living vertebrate neurons. *Journal of Cell Biology*, 131(5):1315–1326. (Cited on page 81.)

Moser, E. I., Kropff, E., and Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, 31:69–89. (Cited on page 10.)

Müller, E. and Thalmann, U. (2000). Origin and evolution of primate social organisation : a reconstruction. *Biological reviews of the Cambridge Philosophical Society*, 75(3):405–35. (Cited on page 79.)

Mullins, R. D., Heuser, J. A., and Pollard, T. D. (1998). The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proceedings of the National Academy of Sciences*, 95(11):6181–6186. (Cited on page 81.)

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, a., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Gene-Wide Identification of Episodic Selection. *Molecular Biology and Evolution*, 32(5):1365–1371. (Cited

on page 36.)

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8(7). (Cited on pages 28, 29, 36, 139, and 159.)

Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Computer Journal*, 26(4):354–359. (Cited on page 57.)

Murtagh, F. and Legendre, P. (2014). Ward ' s Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward ' s Criterion ? *Journal of Classification*, 31(October):274–295. (Cited on page 57.)

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, 11(5):715–24. (Cited on page 28.)

Myrum, C., Baumann, A., Bustad, H., Flydal, M., Mariaule, V., Alvira, S., Cuéllar, J., Haavik, J., Soulé, J., Valpuesta, J., Márquez, J. A., Martinez, A., and Bramham, C. (2015). Arc is a flexible modular protein capable of reversible self-oligomerization. *Biochemical Journal*, pages 145–158. (Cited on pages 131 and 132.)

Nair, R. R., Patil, S., Tiron, A., Kanhema, T., Panja, D., Schiro, L., Parboczak, K., Wilczynski, G., and Clive, R. (2017). Dynamic Arc SUMOylation and selective interaction with F-actin-binding protein drebrin A in LTP consolidation in vivo. *Front Synaptic Neurosci*, 9(May):10.3389/fnsyn.2017.00008. (Cited on pages 134 and 135.)

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*. (Cited on page 21.)

Nei, M. (1972). Genetic Distance between Populations. *The American Naturalist*, 106(949):283–292. (Cited on page 23.)

Nei, M. (1976). Mathematical Models of Speciation and Genetic Distance. (Cited on page 23.)

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford university press. (Cited on pages 16 and 25.)

Neves, S. R., Ram, P. T., and Iyengar, R. (2002). G Protein Pathways. *Science*, 296(5573):1636–1639. (Cited on page 8.)

Newton, A. C. (1995). Protein Kinase C: Structure, Function, and Regulation*. *The Journal of biological chemistry*, 1(27):28495–28498. (Cited on page 89.)

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17. (Cited on pages 22 and 34.)

Novikova, I. V., Hennelly, S. P., and Sanbonmatsu, K. Y. (2012). Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture*, 2(6):189–99. (Cited on page 169.)

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G., and Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):358–363. (Cited on page 91.)

Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L., Danila, M. I., Feng, G., and Chisholm, R. L. (2009). Annotating the human genome with Disease Ontology. *BMC Genomics*, 10(Suppl 1):S6. (Cited on page 45.)

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., and Hofreiter, M. (2004). Genetic Analyses from Ancient DNA. *Annual Review of Genetics*, 38(1):645–679. (Cited on page 1.)

Pai, B., Siripornmongcolchai, T., Berentsen, B., Pakzad, A., Vieuille, C., Pallesen, S., Pajak, M., Simpson, T. I., Armstrong, J. D., Wibrand, K., and Bramham, C. R. (2014). NMDA receptor-dependent regulation of miRNA expression and association with Argonaute during LTP in vivo. *Frontiers in Cellular Neuroscience*, 7(January):1–15. (Cited on page 163.)

Pais, F. S.-M., Ruy, P. D. C., Oliveira, G., and Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9:4. (Cited on page 34.)

Pfeiffer, B. E. and Huber, K. M. (2006). Current Advances in Local Protein Synthesis and Synaptic Plasticity. *Journal of Neuroscience*, 26(27):7147–7150. (Cited on page 10.)

Pin, J. P. (2000). Molecular tinkering of G protein-coupled receptors : an evolutionary success. *EMBO Journal*, 18(7):1723–1729. (Cited on page 7.)

Pinsker, H. M., Hening, W. A., Carew, T. J., and Kandel, E. R. (1973). Long-Term Sensitization of a Defensive Withdrawal Reflex in Aplysia. *Science*, (5):1039–1042. (Cited on page 11.)

Plath, N., Ohana, O., Dammermann, B., Errington, M. L., Schmitz, D., Gross, C., Mao, X., Engelsberg, A., Mahlke, C., Welzl, H., Kobalz, U., Stawrakakis, A., Fernandez, E., Waltereit, R., Bick-Sander, A., Therstappen, E., Cooke, S. F., Blanquet, V., Wurst, W., Salmen, B., Bösl, M. R., Lipp, H.-P., Grant, S. G. N., Bliss, T. V. P., Wolfer, D. P., and Kuhl, D. (2006). Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron*, 52(3):437–44. (Cited on pages 10, 11, and 132.)

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287(1):187–198. (Cited on page 89.)

Pond, S. L. K. and Frost, S. D. W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)*, 21(10):2531–3. (Cited on page 36.)

Pond, S. L. K., Frost, S. D. W., Grossman, Z., Gravenor, M. B., Richman, D. D., and Brown, A. J. L. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS computational biology*, 2(6):e62. (Cited on page 31.)

Posada, D. and Crandall, K. A. (2002). The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal of molecular evolution*, 54(3):396–402. (Cited on page 42.)

Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., Lamantia, A.-S., and White, L. (2012). *Neuroscience*. Sinauer Associates, Inc., Sunderland, MA, 5 edition. (Cited on page 3.)

Raimondi, A., Ferguson, S. M., Lou, X., Armbruster, M., Paradise, S., Giovedi, S., Messa, M., Kono, N., Takasaki, J., Cappello, V., O'Toole, E., Ryan, T. A., and De Camilli, P. (2011). Overlapping Role of Dynamin Isoforms in Synaptic Vesicle Endocytosis. *Neuron*, 70(6):1100–1114. (Cited on page 117.)

Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics*, 10(7). (Cited on pages 2 and 168.)

Renne, P. R., Deino, A. L., Hilgen, F. J., Kuiper, K. F., Mark, D. F., Mitchell, W. S., Morgan, L. E., Mundil, R., and Smit, J. (2013). Time Scales of Critical Events Around the Cretaceous-Paleogene Boundary. *Science*, 339(6120):684–687. (Cited on page 80.)

Richards, G. S., Simionato, E., Perron, M., Adamska, M., Vervoort, M., and Degnan, B. M. (2008). Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. *Current biology : CB*, 18(15):1156–61. (Cited on page 11.)

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics (Oxford, England)*, 30(7):1020–1021. (Cited on

page 31.)

Rodriguez, F., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of theoretical biology*, 142(4):485–501. (Cited on page 24.)

Rogaeva, E. a., Fafel, K. C., Song, Y. Q., Medeiros, H., Sato, C., Liang, Y., Richard, E., Rogaev, E. I., Frommelt, P., Sadovnick, a. D., Meschino, W., Rockwood, K., Boss, M. a., Mayeux, R., and St George-Hyslop, P. (2001). Screening for PS1 mutations in a referral-based series of AD cases: 21 novel mutations. *Neurology*, 57(4):621–5. (Cited on page 156.)

Ross, C. (1996). Adaptive explanation for the origins of the Anthropoidea (primates). *American Journal of Primatology*, 40(3):205–230. (Cited on page 79.)

Ross, H. E., Cole, C. D., Smith, Y., Neumann, I. D., Landgraf, R., Murphy, A. Z., and Young, L. J. (2009). Characterization of the oxytocin system regulating affiliative behavior in female prairie voles. *Neuroscience*, 162(4):892–903. (Cited on page 118.)

Rouquier, S., Blancher, a., and Giorgi, D. (2000). The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 97(6):2870–2874. (Cited on page 79.)

Rumpel, S., Ledoux, J., Zador, A., and Malinow, R. (2005). Postsynaptic Receptor Trafficking Underlying a Form of Associative Learning. *Science*, 308(April):83–89. (Cited on page 10.)

Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F., and Gorodkin, J. (2013). RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. *Human Mutation*, 34(4):546–556. (Cited on page 170.)

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–25. (Cited on pages 22 and 26.)

Sakarya, O., Armstrong, K. a., Adamska, M., Adamski, M., Wang, I.-F., Tidor, B., Degnan, B. M., Oakley, T. H., and Kosik, K. S. (2007). A post-synaptic scaffold at the origin of the animal kingdom. *PloS one*, 2(6):e506. (Cited on page 11.)

Samaco, R. C., Mandel-Brehm, C., Chao, H.-T., Ward, C. S., Fyffe-Maricich, S. L., Ren, J., Hyland, K., Thaller, C., Maricich, S. M., Humphreys, P., Greer, J. J., Percy, A., Glaze, D. G., Zoghbi, H. Y., and Neul, J. L. (2009). Loss of MeCP2 in aminergic neurons causes cell-autonomous defects in neurotransmitter synthesis and specific behavioral abnormalities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21966–71. (Cited on pages 7 and 11.)

Schafer, D. A., Weed, S. A., Binns, D., Karginov, A. V., Parsons, J. T., and Cooper, J. A. (2002). Dynamin2 and cortactin regulate actin assembly and filament organization. *Current Biology*, 12(21):1852–1857. (Cited on page 117.)

Schenker, N. M., Desgouttes, A. M., and Semendeferi, K. (2005). Neural connectivity and cortical substrates of cognition in hominoids. *Journal of Human Evolution*, 49(5):547–569. (Cited on page 12.)

Schierup, M. H. and Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–91. (Cited on page 41.)

Schulte, P., Alegret, L., Arenillas, I., Arz, J. A., Barton, P. J., Bown, P. R., Bralower, T. J., Christeson, G. L., Claeys, P., Cockell, C. S., Collins, G. S., Deutsch, A., Goldin, T. J., Goto, K., Grajales-Nishimura, J. M., Grieve, R. A. F., Gulick, S. P. S., Johnson, K. R., Kiessling, W., Koeberl, C., Kring, D. A., MacLeod, K. G., Matsui, T., Melosh, J., Montanari, A., Morgan, J. V., Neal, C. R., Nichols, D. J., Norris, R. D., Pierazzo, E., Ravizza, G., Rebolledo-Vieyra, M., Reimold, W. U., Robin, E., Salge, T., Speijer, R. P., Sweet, A. R., Urrutia-Fucugauchi, J., Vajda, V., Whalen, M. T., and Willumsen, P. S. (2010). The Chicxulub Asteroid Impact and Mass Extinction at the Cretaceous-Paleogene Boundary. *Science*, 327(5970):1214–1218. (Cited on page 80.)

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*. (Cited on page 24.)

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504. (Cited on pages 102 and 113.)

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular biology and evolution*, 23(1):7–9. (Cited on page 28.)

Shen, S. H., Slightom, J. L., and Smithies, O. (1981). A history of the human fetal globin gene duplication. *Cell*, 26(2 Pt 2):191–203. (Cited on page 41.)

Shepherd, J. D., Rumbaugh, G., Wu, J., Chowdhury, S., Plath, N., Kuhl, D., Huganir, R. L., and Worley, P. F. (2006). Arc/Arg3.1 Mediates Homeostatic Synaptic Scaling of AMPA Receptors. *Neuron*, 52(3):475–484. (Cited on pages 131 and 134.)

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, 32(6):2616–2641. (Cited on page 59.)

Shupliakov, O., Löw, P., Grabs, D., Gad, H., Chen, H., David, C., Takei, K., Camilli, P. D., Brodin, L., Lw, P., and Brodint, L. (1997). Synaptic Vesicle Endocytosis Impaired by Disruption of Dynamin-SH3 Domain Interactions. *Science*, 276(5310):259–263. (Cited on page 6.)

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539). (Cited on pages 22 and 34.)

Sipos, B., Massingham, T., Jordan, G. E., and Goldman, N. (2011). PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC bioinformatics*, 12(1):104. (Cited on page 166.)

Smith, A. (1994). Rooting molecular trees: problems and strategies. *Biological Journal of the Linnean Society*. (Cited on page 27.)

Smith, J. D., McManus, K. F., and Fraser, H. B. (2013). A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. *Molecular Biology and Evolution*, 30(11):2509–2518. (Cited on page 169.)

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Molecular Biology and Evolution*, 32(5):1342–1353. (Cited on pages 29, 30, 53, and 160.)

Soderling, T. R. (2000). CaM-kinases: Modulators of synaptic plasticity. *Current Opinion in Neurobiology*, 10(3):375–380. (Cited on page 9.)

Sokal, R. R. and Sneath, P. H. A. (1963). Principles of numerical taxonomy. *Principles of numerical taxonomy.* (Cited on page 26.)

Somel, M., Sayres, M. A., Jordan, G., Huerta-Sanchez, E., Fumagalli, M., Ferrer-Admetlla, A., and Nielsen, R. (2013). A scan for human-specific relaxation of negative selection reveals unexpected polymorphism in proteasome genes. *Molecular Biology and Evolution*. (Cited on page 118.)

Sorenson, H. W. (1980). *Parameter estimation: principles and problems*. Marcel Dekker New York. (Cited on page 24.)

Strauss, T. and Von Maltitz, M. J. (2017). Generalising ward's method for use with manhattan distances. *PLoS ONE*, 12(1):1–21. (Cited on page 57.)

Strope, C. L., Scott, S. D., and Moriyama, E. N. (2007). Indel-Seq-Gen: A new protein family simulator incorporating domains, motifs, and indels. *Molecular Biology and Evolution*, 24(3):640–649. (Cited on page 166.)

Su, K.-Y., Chien, W.-L., Fu, W.-M., Yu, I.-S., Huang, H.-P., Huang, P.-H., Lin, S.-R., Shih, J.-Y., Lin, Y.-L., Hsueh, Y.-P., Yang, P.-C., and Lin, S.-W. (2007). Mice Deficient in Collapsin Re-

sponse Mediator Protein-1 Exhibit Impaired Long-Term Potentiation and Impaired Spatial Learning and Memory. *Journal of Neuroscience*, 27(10):2513–2524. (Cited on page 81.)

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50. (Cited on page 46.)

Südhof, T. C. (2004). THE SYNAPTIC VESICLE CYCLE. *Annual Review of Neuroscience*, 27(1):509–547. (Cited on page 6.)

Südhof, T. C. (2006). Synaptic Vesicles: An Organelle Comes of Age. *Cell*, 127(4):671–673. (Cited on page 6.)

Sullivan, P. F., Keefe, R. S. E., Lange, L. A., Lange, E. M., Stroup, T. S., Lieberman, J., and Maness, P. F. (2007). NCAM1 and Neurocognition in Schizophrenia. *Biological Psychiatry*, 61(7):902–910. (Cited on page 81.)

Suzuki, R. and Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542. (Cited on page 59.)

Takamori, S., Holt, M., Stenius, K., Lemke, E. A., Grønborg, M., Riedel, D., Urlaub, H., Schenck, S., Brügger, B., Ringler, P., Müller, S. A., Rammner, B., Gräter, F., Hub, J. S., De Groot, B. L., Mieskes, G., Moriyama, Y., Klingauf, J., Grubmüller, H., Heuser, J., Wieland, F., and Jahn, R. (2006). Molecular Anatomy of a Trafficking Organelle. *Cell*, 127(4):831–846. (Cited on page 6.)

Talavera, D., Lovell, S. C., and Whelan, S. (2015). Covariation is a poor measure of molecular coevolution. *Molecular Biology and Evolution*, 32(9):2456–2468. (Cited on page 89.)

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4):564–77. (Cited on pages 152 and 165.)

Tamayo, P., Steinhardt, G., Liberzon, A., and Mesirov, J. P. (2016). The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):472–487. (Cited on page 82.)

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–26. (Cited on page 24.)

Tatusova, T. a. and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters*, 174(2):247–50. (Cited on page 19.)

Thivierge, J. P. and Marcus, G. F. (2007). The topographic brain: from neural connectivity to cognition. *Trends in Neurosciences*, 30(6):251–259. (Cited on page 12.)

Thomas, G. M. and Huganir, R. L. (2004). MAPK cascade signalling and synaptic plasticity. *Nature Reviews Neuroscience*, 5(3):173–183. (Cited on page 8.)

Thomas, P. D., Campbell, M., Kejariwal, A., Mi, H., and Karlak, B. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, 13(9):2129–2141. (Cited on pages 45 and 76.)

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–80. (Cited on page 22.)

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–36. (Cited on page 22.)

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(05):675–735. (Cited on page 1.)

Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Research*, 16(7):885–889. (Cited on page 169.)

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419. (Cited on page 40.)

Urban, N. N. and Barrionuevo, G. (1996). Induction of hebbian and non-hebbian mossy fiber long-term potentiation by distinct patterns of high-frequency stimulation. *The Journal of neuroscience*, 16(13):4293–9. (Cited on page 8.)

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19:327–335. (Cited on pages 20 and 33.)

Villar, D., Berthelot, C., Flicek, P., Odom, D. T., Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., and Pignatelli, M. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566. (Cited on page 163.)

Volders, P.-J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., and Mestdagh, P. (2014). An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research*, 43(D1):D174–D180. (Cited on page 170.)

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403. (Cited on pages 91, 120, and 121.)

Waites, C. L., Craig, A. M., and Garner, C. C. (2005). Mechanisms of Vertebrate Synaptogenesis. *Annual review of neuroscience*, pages 251–276. (Cited on page 5.)

Waltereit, R. and Weller, M. (2003). Signaling from cAMP/PKA to MAPK and synaptic plasticity. *Molecular neurobiology*, 27(1):99–106. (Cited on pages 8 and 9.)

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. (Cited on page 57.)

Wool, I. G. (1996). Extraribosomal functions of ribosomal proteins. *Trends in Biochemical Sciences*, 21(5):164–165. (Cited on page 81.)

Wool, I. G., Chan, Y.-L., and Glück, A. (1995). Structure and evolution of mammalian ribosomal proteins. *Biochemistry and Cell Biology*, 73(11-12):933–947. (Cited on page 81.)

Wu, G., Dawson, E., Duong, A., Haw, R., and Stein, L. (2014). ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research*, pages 1–14. (Cited on pages 102 and 113.)

Wu, J., Petralia, R. S., Kurushima, H., Patel, H., Jung, M.-y., Volk, L., Chowdhury, S., Shepherd, J. D., Dehoff, M., Li, Y., Kuhl, D., Huganir, R. L., Price, D. L., Scannevin, R., Troncoso, J. C., Wong, P. C., and Worley, P. F. (2011). Arc / Arg3 . 1 Regulates an Endosomal Pathway Essential for Activity-Dependent b -Amyloid Generation. *Cell*, 147(3):615–628. (Cited on pages 132 and 135.)

Xenarios, I. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305. (Cited on page 91.)

Xu, Z. and Mathews, D. H. (2011). Multilign: An algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics*, 27(5):626–632. (Cited on page 170.)

Yamashita, N., Uchida, Y., Ohshima, T., Hirai, S.-i., Nakamura, F., Taniguchi, M., Mikoshiba, K., Honnorat, J., Kolattukudy, P., Thomasset, N., Takei, K., Takahashi, T., and Goshima, Y. (2006). Collapsin response mediator protein 1 mediates reelin signaling in cortical neuronal migration. *J. Neurosci.*, 26(51):13357–13362. (Cited on page 81.)

Yang, B. Z., Kranzler, H. R., Zhao, H., Gruen, J. R., Luo, X., and Gelernter, J. (2008a). Haplotypic variants in DRD2, ANKK1, TTC12, and NCAM1 are associated with comorbid alcohol and drug dependence. *Alcoholism: Clinical and Experimental Research*, 32(12):2117–2127. (Cited on page 81.)

Yang, H.-P., Wang, L., Han, L., and Wang, S. C. (2013). Nonsocial functions of hypothalamic oxytocin. *ISRN neuroscience*, 2013:179272. (Cited on page 118.)

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–6. (Cited on page 155.)

Yang, W., Steen, H., and Freeman, M. R. (2008b). Proteomic approaches to the analysis of multiprotein signaling complexes. *Proteomics*, 8(4):832–851. (Cited on page 89.)

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556. (Cited on pages 27 and 36.)

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91. (Cited on pages 27 and 36.)

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports*, 6(July):30750. (Cited on page 87.)

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, a. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49. (Cited on page 28.)

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196. (Cited on page 48.)

Yoshimura, Y., Yamauchi, Y., Shinkawa, T., Taoka, M., Donai, H., Takahashi, N., Isobe, T., and Yamauchi, T. (2004). Molecular constituents of the postsynaptic density fraction revealed by proteomic analysis using multidimensional liquid chromatography-tandem mass spectrometry. *Journal of Neurochemistry*, 88(3):759–768. (Cited on page 44.)

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298. (Cited on page 41.)

Zhang, W., Wu, J., Ward, M., Yang, S., Chuang, Y.-A., Xiao, M., Li, R., Leahy, D., and Worley, P. (2015). Structural Basis of Arc Binding to Synaptic Proteins: Implications for Cognitive Disease. *Neuron*, pages 1–11. (Cited on pages 131, 132, and 134.)

Zhao, S. and Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):1–14. (Cited on page 33.)

Zucker, R. S. and Regehr, W. G. (2002). Short-Term Synaptic Plasticity. *Annual Review of Physiology*, 64(1):355–405. (Cited on page 8.)

A

SUPPLEMENTARY TABLES AND FIGURES

IMPLEMENTATION OF MODELLING WORKFLOW AND ANALYSIS TOOLS

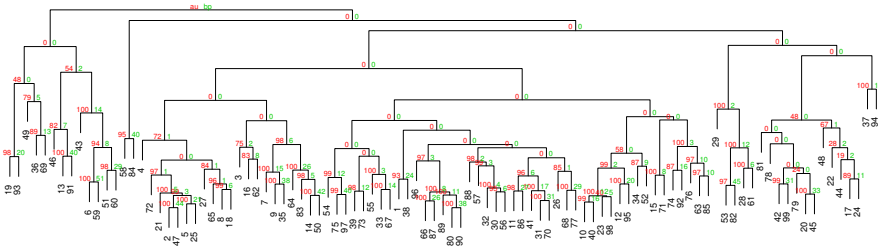
Source code, as well as aggregated modelling results in the form of SQLite database files, are attached with the electronic copy of the examination version of the thesis. Source code and auxiliary files, which would allow for replication of the entire study are also maintained in a private `git` repository.
Further to that, `Hyphy` source is available from the github repository (github.com/veg/hyphy). `Clustal Omega` source is available from authors' website (www.clustal.org/omega). `OpenMPI` has been maintained on Eddie supercomputer by support service but it also available from project's website (www.open-mpi.org). Historical versions of Ensembl Compara sequence data are easily accessible through `biomaRt` package in `R`.
`R` packages which cannot be loaded automatically from `CRAN/Bioconductor` are:

- `topOnto` used for multi-ontology enrichement tests (github.com/hxin/topOnto)
- `ograph` used to ontology graph manipulation (github.com/hxin/ograph)
- `igraph` in version 0.7.1 as a requirement for `ograph` - historical version needs to be downloaded manually from `CRAN` (up-to-date version of `igraph` is used elsewhere).

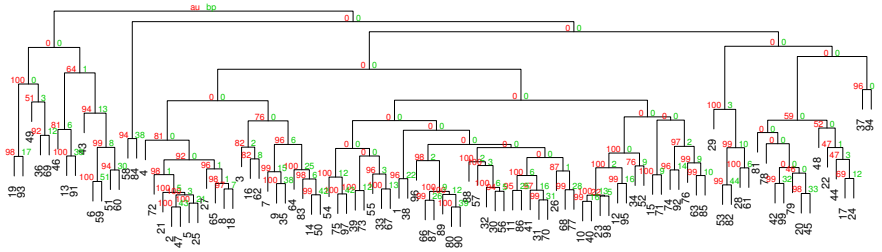
All other `R` packages used are available on `CRAN/Bioconductor` as of July 2017. Similarly, `biopython` is available through `pip` package management system for `python`. Finally, `R`, `python`, or `gcc` version should not have effect on replicability of results (might marginally affect speed of execution), up-to-date versions have been maintained throughout the lifespan of the project.

Table A.1: Source code scripts list.

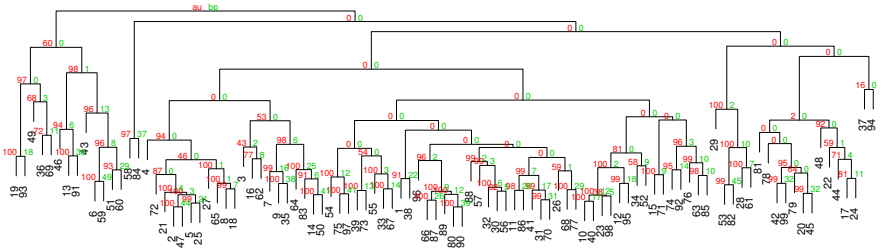
Task	source files
Modelling setup	`latex_tabling.R, order.R, file_operations.R, full_proteome_prep.R, sorting_species.R, removed_species_analysis.R, remove_species_from_fastas.R, tree_making.R, protein_group_extraction.R, obtain_ensembl_orthos.R, save_fastas_with_orthos.R, pre_alignment_controller.R`
Alignment and selection inference jobs	`preopenmp_dtc.sh, preopenmpi_scatter_extendedtwice.sh, python_phylo_job_template_short.sh, align_jobs_generate.R, helpers.py, transcript_ali.py, compare_alignments.R, diagnostics_missing_files.R, hyphy_commands_generator.R`
Modelling results collection	`absrel_postanalysis.R, meme_postanalysis.R, meme_site_postanalysis.R, post_analysis_helpers.R, post_hyphy_universal.R`
Chapter 3	`c_functions.R, plotBranchbyTrait.R, absrel_post_postanalysis.R, enrichments_clusters.R, stats_extraction.R, psp_cluster_crosstabs.R`
Chapter 4	`psp_network.R, psp_communities.R, psp_equivalence.R, psp_radiation_paths.R, psp_radiation.R, psp_reactome.R, psp_size_analysis.R`
Chapter 5	`arc_genes_sorting.R, domain_search_universal.R, phospho_aggregation.R, arcome_results_comparison.R, arc_stuff.R, per_gene_summaries.R`



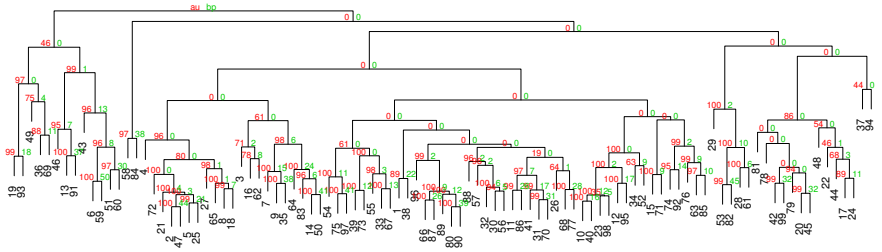
(a) $N_{bootstraps} = 250$



(b) $N_{bootstraps} = 500$

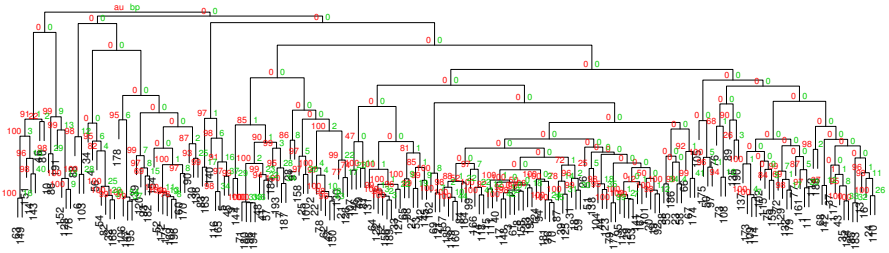


(c) $N_{bootstraps} = 1000$

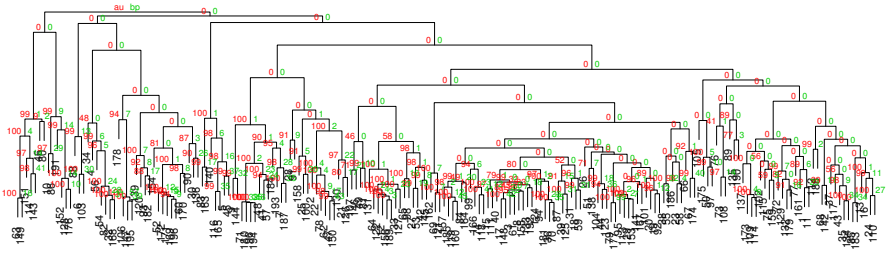


(d) $N_{bootstraps} = 2000$

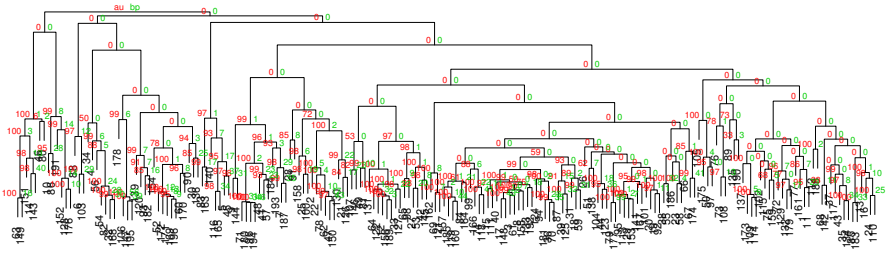
Figure A.1: Bootstrap clustering tests for a sample of 100 proteins.



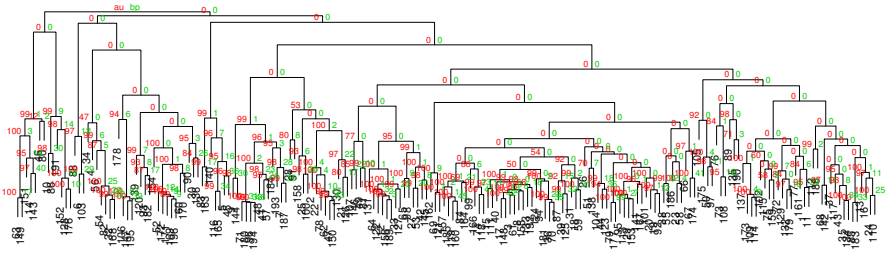
(a) $N_{bootstraps} = 250$



(b) $N_{bootstraps} = 500$



(c) $N_{bootstraps} = 1000$



(d) $N_{bootstraps} = 2000$

Figure A.2: Bootstrap clustering tests for a sample of 200 proteins.

Table A.2: Selection of the cluster number for the complete linkage method. K is the number of clusters in the tree cut, then measures of intra-cluster distance follow, max(max) refers to maximum of all maximum distances in each of the K clusters. min(median) is the lowest out of all median distances in each of the K clusters, respectively max is the highest one. min(n_i) is the size of the smallest cluster, and respectively max(n_i) is the size of the largest one. Compare Table 3.3 for similar statistics for a different method of hierarchical clustering.

K	Intra distance: max(max)	min(median)	max(median)	min(n_i)	max(n_i)
5	1.999	0.758	1.279	1217	6776
6	1.999	0.544	1.279	1217	6776
7	1.997	0.544	1.199	540	6776
8	1.988	0.544	1.082	215	6776
9	1.971	0.544	1.082	215	3607
10	1.965	0.544	1.055	215	3607
11	1.933	0.544	1.055	215	3607
12	1.927	0.518	1.055	137	3470
13	1.914	0.518	1.031	137	3470
14	1.891	0.518	0.883	137	3470
15	1.887	0.463	0.883	137	3470
16	1.865	0.463	0.883	137	3470
17	1.858	0.409	0.883	137	3470
18	1.85	0.409	0.883	137	3470
19	1.837	0.409	0.883	77	3470
20	1.834	0.409	0.883	77	3470

Table A.3: Full names of all taxa, sorted alphabetically.

Full name	Abbreviated name	Name in figures
Ailuropoda melanoleuca	*A. melanoleuca*	AMELANOLEU
Anas platyrhynchos	*A. platyrhynchos*	APLATYRHYN
Anolis carolinensis	*A. carolinensis*	ACAROLINEN
Astyanax mexicanus	*A. mexicanus*	AMEXICANUS
Bos taurus	*B. taurus*	BTAURUS
Caenorhabditis elegans	*C. elegans*	CELEGANS
Callithrix jacchus	*C. jacchus*	CJACCHUS
Canis familiaris	*C. familiaris*	CFAMILIARI

continued ...

. . . continued

Full name	Abbreviated name	Name in figures
Carlito syrichta	*T. syrichta*	TSYRICHTA
Cavia porcellus	*C. porcellus*	CPORCELLUS
Chlorocebus sabaeus	*C. sabaeus*	CSABAEUS
Choloepus hoffmanni	*C. hoffmanni*	CHOFFMANNI
Ciona intestinalis	*C. intestinalis*	CINTESTINA
Ciona savignyi	*C. savignyi*	CSAVIGNYI
Danio rerio	*D. rerio*	DRERIO
Dasypus novemcinctus	*D. novemcinctus*	DNOVEMCINC
Dichotomyctere nigroviridis	*T. nigroviridis*	TNIGROVIRI
Dipodomys ordii	*D. ordii*	DORDII
Drosophila melanogaster	*D. melanogaster*	DMELANOGAS
Echinops telfairi	*E. telfairi*	ETELFAIRI
Equus caballus	*E. caballus*	ECABALLUS
Erinaceus europaeus	*E. europaeus*	EEUROPAEUS
Felis catus	*F. catus*	FCATUS
Ficedula albicollis	*F. albicollis*	FALBICOLLI
Gadus morhua	*G. morhua*	GMORHUA
Gallus gallus	*G. gallus*	GGALLUS
Gasterosteus aculeatus	*G. aculeatus*	GACULEATUS
Gorilla gorilla	*G. gorilla*	GGORILLA
Homo sapiens	*H. Sapiens*	HSAPIENS
Ictidomys tridecemlineatus	*I. tridecemlineatus*	ITRIDECEML
Latimeria chalumnae	*L. chalumnae*	LCHALUMNAE
Lepisosteu oculatus	*L. oculatus*	LOCULATUS
Loxodonta africana	*L. africana*	LAFRICANA
Macaca mulatta	*M. mulatta*	MMULATTA
Macropus eugenii	*M. eugenii*	MEUGENII
Meleagris gallopavo	*M. gallopavo*	MGALLOPAVO

continued . . .

. . . continued

Full name	Abbreviated name	Name in figures
Microcebus murinus	*M. murinus*	MMURINUS
Monodelphis domestica	*M. domestica*	MDOMESTICA
Mus musculus	*M. musculus*	MMUSCULUS
Mustela putorius furo	*M. furo*	MFURO
Myotis lucifugus	*M. lucifugus*	MLUCIFUGUS
Nomascus leucogenys	*N. leucogenys*	NLEUCOGENY
Ochotona princeps	*O. princeps*	OPRINCEPS
Oreochromis niloticus	*O. niloticus*	ONILOTICUS
Ornithorhynchus anatinus	*O. anatinus*	OANATINUS
Oryctolagus cuniculus	*O. cuniculus*	OCUNICULUS
Oryzias latipes	*O. latipes*	OLATIPES
Otolemu garnettii	*O. garnettii*	OGARNETTII
Ovis aries	*O. aries*	OARIES
Pan troglodytes	*P. troglodytes*	PTROGLODYT
Papio anubis	*P. anubis*	PANUBIS
Pedetes capensis	*P. capensis*	PCAPENSIS
Pelodiscus sinensis	*P. sinensis*	PSINENSIS
Petromyzon marinus	*P. marinus*	PMARINUS
Poecilia formosa	*P. formosa*	PFORMOSA
Pongo abelii	*P. abelii*	PABELII
Pteropus vampyrus	*P. vampyrus*	PVAMPYRUS
Rattus norvegicus	*R. norvegicus*	RNORVEGICU
Saccharomyces cerevisiae	*S. cerevisiae*	SCEREVISIA
Sarcophilus harrisii	*S. harrisii*	SHARRISII
Sorex araneus	*S. araneus*	SARANEUS
Sus scrofa	*S. scrofa*	SSCROFA
Taeniopygia guttata	*T. guttata*	TGUTTATA
Takifugu rubripes	*T. rubripes*	TRUBRIPES

continued . . .

. . . continued

Full name	Abbreviated name	Name in figures
Tupaia belangeri	*T. belangeri*	TBELANGERI
Tursiops truncatus	*T. truncatus*	TTRUNCATUS
Vicugna pacos	*V. pacos*	VPACOS
Xenopus tropicalis	*X. tropicalis*	XTROPICALI
Xiphophorus maculatus	*X. maculatus*	XMACULATUS

Table A.4: Average sequence similarity of ortholog sequences to reference human sequence in the set of transcripts of *Arc* interactome proteins.

Taxon	Similarity (%)				
scerevisiae	36.0	xtropicalis	72.1	ttruncatus	86.8
celegans	43.8	acarolinensis	72.1	mmulatta	87
csavignyi	44.5	aplatyrhynchos	72.6	oaries	87.5
cintestinalis	45.7	tsyrichta	72.8	ocuniculus	88
pmarinus	47.1	vpacos	73.5	itridecemlineatus	88
dmelanogaster	49.0	tbelangeri	73.5	ecaballus	88.6
oanatinus	60.7	meugenii	73.8	cfamiliaris	89
saraneus	63.2	mgallopavo	74.1	mfuro	89.7
gmorhua	64.0	eeuropaeus	75.4	pvampyrus	89.9
olatipes	66.1	falbicollis	75.8	fcatus	89.9
drerio	66.6	ggallus	77.2	panubis	90.8
amexicanus	66.8	psinensis	78.2	btaurus	91.2
lchalumnae	66.9	mmurinus	78.5	ggorilla	91.5
pformosa	67.4	dordii	78.7	mmusculus	91.8
xmaculatus	67.6	oprinceps	79.3	amelanoleuca	92
tnigroviridis	67.7	sscrofa	80.1	cporcellus	92.1
gaculeatus	69.6	sharrisii	80.2	nleucogenys	92.3
tguttata	69.8	pcapensis	81.9	pabelii	92.8
choffmanni	70.2	mdomestica	82.9	rnorvegicus	93.7
trubripes	70.6	lafricana	84.8	cjacchus	93.7
loculatus	70.9	dnovemcinctus	85.6	csabaeus	94.8
oniloticus	71.1	mlucifugus	85.6	ptroglodytes	95
etelfairi	71.4	ogarnettii	86.5		

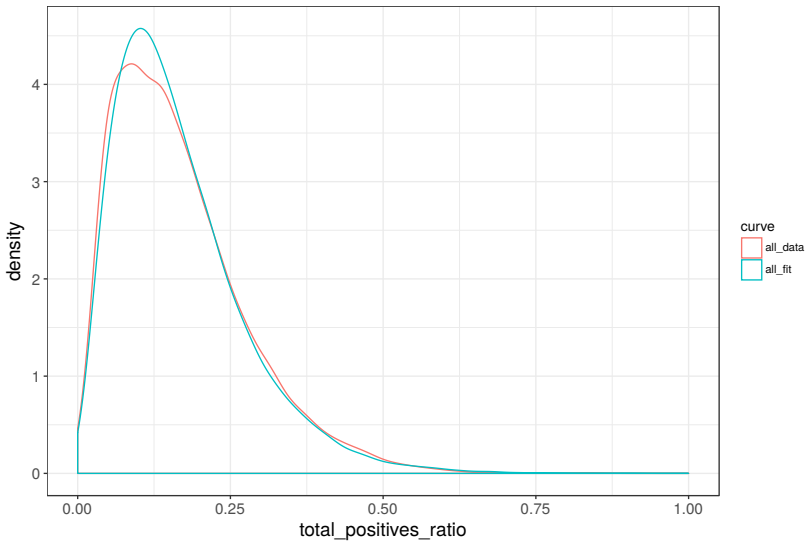


Figure A.3: Gamma distribution fit to the relative total number of positives, parameters derived from data through the method of moments.

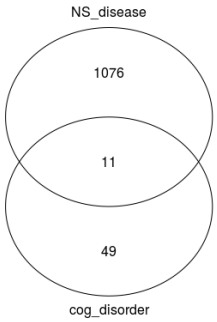


Figure A.4: Overlap of terms picked for the reduced Human Disease Ontology, NS_disease is a set of all descendants of a high level node - *Nervous System Disease*, cog_disorder is a set of all descendants of equally high level node of *Cognitive disorder*

Table A.6: Comparison of alignmnent lengths between full and reduced tree for Arc complex proteins; deepest ancestral node in the full phylogenetic tree, consult table 3.2 for dating, node 23 indicates protein ortholog is present in yeast which was used to root the tree

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
M6PR	21	491	481	2.04%	57	43	14

continued ...

. . . continued

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
COX15	23	520	423	18.65%	57	42	15
VDAC3	23	327	310	5.2%	61	45	16
HSDL1	23	424	394	7.08%	59	45	14
TVP23C	23	335	318	5.07%	49	27	22
C1QBP	22	372	350	5.91%	57	40	17
CCDC47	22	550	510	7.27%	63	46	17
EFNB2	22	710	365	48.59%	61	45	16
PTGER2	21	440	416	5.45%	56	42	14
MRPL15	23	378	344	8.99%	64	47	17
FXYD6	19	275	271	1.45%	45	36	9
RPLP1	22	125	118	5.6%	51	40	11
NOL4	22	952	749	21.32%	42	39	3
ARL1	23	204	198	2.94%	62	44	18
CREBBP	23	3635	2731	24.87%	51	36	15
GNAO1	23	462	362	21.65%	57	43	14
AP2S1	23	316	303	4.11%	53	37	16
NOTCH3	22	3044	2666	12.42%	41	27	14
SDHA	23	758	736	2.9%	58	41	17
GOLPH3	23	360	317	11.94%	62	44	18
TNPO3	23	1136	1005	11.53%	55	37	18
EIF2B2	23	451	409	9.31%	65	47	18
RTN1	22	1043	909	12.85%	54	37	17
PML	19	1316	1246	5.32%	34	34	0
AKAP8	19	884	861	2.6%	39	37	2
SH3GL1	22	450	421	6.44%	56	40	16
CERS2	22	475	456	4%	56	41	15
C1orf35	20	354	326	7.91%	48	36	12
NOTCH1	22	2987	2655	11.11%	45	30	15
SEC24D	23	1403	1149	18.1%	52	37	15
PRKAG2	23	822	759	7.66%	58	40	18
SLC39A14	23	752	657	12.63%	57	41	16

continued . . .

. . . continued

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
UBE2Q1	22	491	434	11.61%	41	35	6
ATP5J2	22	526	105	80.04%	52	38	14
AP2M1	23	535	495	7.48%	60	43	17
TUBA1A	23	494	486	1.62%	38	34	4
CEP44	20	504	489	2.98%	50	42	8
CACNG2	19	476	333	30.04%	52	42	10
CCNDBP1	19	552	540	2.17%	50	41	9
ITFG3	19	684	774	-13.16%	50	38	12
TADA3	22	598	453	24.25%	57	41	16
SCAMP3	22	421	368	12.59%	51	36	15
LETM1	23	1228	848	30.94%	52	38	14
SEMA4C	20	993	928	6.55%	46	35	11
TVP23B	23	259	322	-24.32%	42	30	12
HPCAL1	23	200	199	0.5%	58	41	17
USMG5	19	158	158	0%	48	43	5
GNAI2	23	463	424	8.42%	55	39	16
DLGAP1	22	1360	1104	18.82%	51	39	12
DNAJC14	23	1203	764	36.49%	47	37	10
TUBB6	23	485	480	1.03%	55	42	13
KRBA1	14	1400	1400	0%	29	29	0
SLC25A19	23	419	396	5.49%	58	42	16
NAP1L5	13	199	199	0%	25	25	0
PSEN1	22	712	632	11.24%	54	40	14
GRIN2A	22	1802	1524	15.43%	53	37	16
SRPR	23	741	668	9.85%	58	41	17
METTL7A	19	253	250	1.19%	57	46	11
VDAC2	23	370	361	2.43%	63	44	19
RHOT1	23	840	768	8.57%	54	37	17
THEM6	22	285	238	16.49%	41	27	14
SEC62	23	675	629	6.81%	58	42	16
ABHD12	22	488	415	14.96%	52	37	15

continued . . .

. . . continued

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
HM13	23	689	500	27.43%	58	40	18
ARMCX3	20	408	409	-0.25%	59	46	13
IKBIP	19	583	455	21.96%	56	46	10
ASH2L	22	716	649	9.36%	56	39	17
USP7	23	1512	1145	24.27%	60	42	18
ERGIC3	23	463	421	9.07%	55	41	14
CAMK2A	23	797	769	3.51%	54	40	14
SLC25A10	23	477	459	3.77%	39	23	16
SPTBN4	22	3044	2846	6.5%	40	25	15
DNM2	23	1149	1155	-0.52%	50	36	14
ARC	17	452	452	0%	29	29	0
RAB18	23	408	391	4.17%	59	43	16
WASF1	22	693	615	11.26%	55	41	14
NOL4L	22	975	733	24.82%	54	42	12
MAP2	22	2692	2224	17.38%	52	38	14
OSTC	22	257	253	1.56%	60	44	16
SCAF8	22	2077	1582	23.83%	54	39	15
SEC63	23	888	771	13.18%	54	38	16
RNGTT	22	703	640	8.96%	53	38	15
CNNM2	23	1115	952	14.62%	50	36	14
ARFGAP1	23	606	451	25.58%	54	36	18
LPHN2	22	2199	1535	30.2%	52	38	14
TM9SF3	23	822	674	18%	59	41	18
TMEM35	22	196	178	9.18%	56	43	13
SGPL1	23	659	628	4.7%	57	39	18
RBBP4	23	452	443	1.99%	56	40	16
RXRB	22	740	618	16.49%	48	34	14
TM9SF2	23	838	779	7.04%	56	39	17
KAT5	23	737	641	13.03%	52	37	15
CNNM4	23	1060	841	20.66%	55	36	19
RTN3	22	1354	1294	4.43%	51	36	15

continued . . .

. . . continued

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
PPP2R2A	23	645	531	17.67%	52	38	14
CD99	16	220	220	0%	25	25	0
C5orf51	19	332	330	0.6%	50	38	12
ZMYM2	22	1963	1560	20.53%	41	35	6
CRELD1	22	591	542	8.29%	52	37	15
STT3A	23	1086	1011	6.91%	55	38	17
PTRH2	23	346	199	42.49%	60	44	16
NME1	23	271	297	-9.59%	50	35	15
ERGIC1	23	515	385	25.24%	59	43	16
PPP3CA	23	677	562	16.99%	56	41	15
PPM1A	23	598	477	20.23%	63	46	17
CAMK2B	23	982	794	19.14%	46	33	13
UBE3A	23	1150	934	18.78%	57	41	16
SLC38A1	23	627	552	11.96%	46	41	5
DLG4	19	1014	942	7.1%	42	31	11
VAPA	23	318	304	4.4%	60	44	16
RAP1GDS1	22	697	648	7.03%	51	40	11
RNF216	19	1079	1005	6.86%	45	36	9
TNPO2	23	1258	1206	4.13%	52	36	16
SLC2A1	23	576	519	9.9%	54	37	17
SH3GL3	22	434	418	3.69%	53	38	15
MAVS	19	858	832	3.03%	48	39	9
VEZT	21	960	870	9.38%	51	37	14
GNA13	23	603	382	36.65%	63	46	17
SMIM1	19	81	81	0%	23	19	4
AP2A2	23	1200	1073	10.58%	49	36	13
DNAJC7	23	878	605	31.09%	56	40	16
YIPF6	22	339	313	7.67%	55	40	15
SRPRB	23	367	274	25.34%	45	29	16
RXRA	22	697	588	15.64%	51	35	16
C10orf88	19	527	506	3.98%	55	45	10

continued . . .

... continued

Protein	Origin	Alignment: full	reduced	drop	Orthologs: all	reduced	drop
COG4	23	991	859	13.32%	58	43	15
LAMTOR3	22	140	128	8.57%	56	44	12
CYB5B	23	184	160	13.04%	59	42	17
RNF170	22	701	666	4.99%	53	39	14
SLC1A5	22	694	597	13.98%	52	39	13
PRKAG1	23	711	371	47.82%	57	39	18
SLC35E1	22	437	436	0.23%	39	39	0
GRIN2B	22	1973	1596	19.11%	54	39	15
AATF	23	738	679	7.99%	61	43	18
AP2B1	23	1060	990	6.6%	52	36	16

Table A.7: Summary of phylogenetic analysis results for Arc interactome proteins, sites under positive selection p<0.05 according to likelihood ratio test for three different methods - aBSREL, FEL, and MEME.

Protein	FEL +sites	MEME +sites	aBSREL +branches	Most recent positive(mya)
AATF	2	59	5	Catarrhini(29)
ABHD12	1	32	4	Similiformes(42.6)
AKAP8	2	44	0	
AP2A2	1	64	3	Euarchontoglires(92.3)
AP2B1	0	39	0	
AP2M1	0	22	0	
AP2S1	0	1	1	Haplorrhini(65.2)
ARC	0	4	1	Euarchontoglires(92.3)
ARFGAP1	0	17	4	Similiformes(42.6)
ARL1	0	5	1	Eutheria(104.2)
ARMCX3	0	10	1	Eutheria(104.2)
ASH2L	0	55	1	Eutheria(104.2)
ATP5J2	0	6	2	Hominoidea(20.4)
C10orf88	5	40	2	Eutheria(104.2)
C1orf35	0	17	2	Euarchontoglires(92.3)
C1QBP	0	18	1	Eutheria(104.2)

continued ...

. . . continued

Protein	FEL +sites	MEME +sites	aBSREL +branches	Most recent positive(mya)
C5orf51	1	16	2	Euarchontoglires(92.3)
CACNG2	0	28	0	
CAMK2A	0	39	1	Eutheria(104.2)
CAMK2B	0	56	0	
CCDC47	0	28	0	
CCNDBP1	1	26	3	Eutheria(104.2)
CD99	0	17	1	Mammalia(167.4)
CEP44	3	35	5	Euarchontoglires(92.3)
CERS2	2	10	1	Euarchontoglires(92.3)
CNNM2	3	58	3	Homo Sapiens(0)
CNNM4	4	67	4	Similiformes(42.6)
COG4	1	19	3	Homo Sapiens(0)
COX15	0	26	0	
CREBBP	3	114	4	Similiformes(42.6)
CRELD1	0	40	4	Similiformes(42.6)
CYB5B	1	18	2	Eutheria(104.2)
DLG4	4	41	3	Euarchontoglires(92.3)
DLGAP1	2	73	0	
DNAJC14	2	39	1	Mammalia(167.4)
DNAJC7	1	35	0	
DNM2	2	57	2	Homo Sapiens(0)
EFNB2	1	17	1	Catarrhini(29)
EIF2B2	0	23	2	Euarchontoglires(92.3)
ERGIC1	1	5	0	
ERGIC3	0	13	2	Hominidae(8.8)
FXYD6	0	6	0	
GNA13	0	17	1	Catarrhini(29)
GNAI2	0	12	1	Eutheria(104.2)
GNAO1	4	25	0	
GOLPH3	0	18	0	
GRIN2A	1	64	0	

continued . . .

. . . continued

Protein	FEL +sites	MEME +sites	aBSREL +branches	Most recent positive(mya)
GRIN2B	3	51	4	Primates(74)
HM13	5	38	2	Boreoeutheria(100)
HPCAL1	0	1	0	
HSDL1	0	13	1	Eutheria(104.2)
IKBIP	9	42	2	Homininae(8.8)
ITFG3	2	48	1	Primates(74)
KAT5	1	12	1	Eutheria(104.2)
KRBA1	13	73	1	Euarchontoglires(92.3)
LAMTOR3	0	3	0	
LETM1	3	53	2	Homininae(8.8)
LPHN2	6	97	1	Theria(162.6)
M6PR	0	9	2	Similiformes(42.6)
MAP2	17	108	4	Similiformes(42.6)
MAVS	7	79	4	Similiformes(42.6)
METTL7A	2	14	0	
MRPL15	2	15	1	Mammalia(167.4)
NAP1L5	1	5	1	Euarchontoglires(92.3)
NME1	1	2	2	Hominoidea(20.4)
NOL4	3	43	0	
NOL4L	0	49	2	Theria(162.6)
NOTCH1	8	0	5	Hominidae(15.7)
NOTCH3	5	201	4	Euarchontoglires(92.3)
OSTC	0	9	0	
PML	10	77	2	Theria(162.6)
PPM1A	1	18	1	Similiformes(42.6)
PPP2R2A	0	31	0	
PPP3CA	0	31	0	
PRKAG1	3	15	2	Eutheria(104.2)
PRKAG2	2	57	1	Theria(162.6)
PSEN1	0	42	4	Euarchontoglires(92.3)
PTGER2	2	18	1	Mammalia(167.4)

continued . . .

. . . continued

Protein	FEL +sites	MEME +sites	aBSREL +branches	Most recent positive(mya)
PTRH2	0	5	2	Euarchontoglires(92.3)
RAB18	1	8	1	Eutheria(104.2)
RAP1GDS1	0	15	0	
RBBP4	0	29	0	
RHOT1	2	41	2	Homo Sapiens(0)
RNF170	1	8	0	
RNF216	6	50	3	Hominidae(15.7)
RNGTT	1	41	1	Mammalia(167.4)
RPLP1	0	3	0	
RTN1	4	64	3	Hominoidea(20.4)
RTN3	21	131	4	Catarrhini(29)
RXRA	0	27	2	Primates(74)
RXRB	0	16	1	Eutheria(104.2)
SCAF8	6	61	1	Theria(162.6)
SCAMP3	0	27	0	
SDHA	0	51	4	Primates(74)
SEC24D	2	71	2	Euarchontoglires(92.3)
SEC62	0	27	0	
SEC63	3	91	0	
SEMA4C	2	27	3	Boreoeutheria(100)
SGPL1	3	39	2	Eutheria(104.2)
SH3GL1	0	32	1	Mammalia(167.4)
SH3GL3	3	23	1	Eutheria(104.2)
SLC1A5	3	68	6	Primates(74)
SLC25A10	0	22	3	Eutheria(104.2)
SLC25A19	1	17	1	Euarchontoglires(92.3)
SLC2A1	2	24	3	Eutheria(104.2)
SLC35E1	0	19	0	
SLC38A1	3	44	3	Primates(74)
SLC39A14	3	39	4	Homo Sapiens(0)
SMIM1	0	1	0	

continued . . .

. . . continued

Protein	FEL +sites	MEME +sites	aBSREL +branches	Most recent positive(mya)
SPTBN4	0	137	3	Eutheria(104.2)
SRPR	3	34	0	
SRPRB	0	11	3	Similiformes(42.6)
STT3A	0	18	1	Eutheria(104.2)
TADA3	0	6	1	Eutheria(104.2)
THEM6	0	4	3	Similiformes(42.6)
TM9SF2 8	4	30	4	Catarrhini(29)
TM9SF3	2	30	1	Similiformes(42.6)
TMEM35	0	4	0	
TNPO2	0	39	0	
TNPO3	0	47	1	Eutheria(104.2)
TUBA1A	0	2	0	
TUBB6	0	15	0	
TVP23B	0	15	1	Mammalia(167.4)
TVP23C	0	10	5	Similiformes(42.6)
UBE2Q1	2	29	1	Boreoeutheria(100)
UBE3A	1	38	4	Primates(74)
USMG5	0	1	0	
USP7	9	71	4	Homo Sapiens(0)
VAPA	3	28	3	Hominidae(15.7)
VDAC2	1	29	2	Eutheria(104.2)
VDAC3	1	11	1	Theria(162.6)
VEZT	5	55	3	Eutheria(104.2)
WASF1	0	25	0	
YIPF6	0	20	0	
ZMYM2	0	67	4	Similiformes(42.6)

Table A.5: Mapping nodes of the human path to the timeline of divergence (from Ensembl Compara) in optional analysis with a reduced tree, multiple nodes with the same divergence time and same taxonomic name arise from resolution of amigious trichtomies in the tree.

Node distance	Scientific name	Ensembl name	Divergence time (mya)
1	Homo Sapiens	Human	0.0
2	Homininae	Hominines	8.8
3	Homininae	Hominines	8.8
4	Hominidae	Great Apes	15.7
5	Hominoidea	Apes	20.4
6	Catarrhini	Apes&OW monkeys	29.0
7	Similiformes	Simians	42.6
8	Haplorrhini	Dry-nosed primates	65.2
9	Primates	Primates	74.0
10	Euarchontoglires	Primates&Rodents	92.3
11	Euarchontoglires	Primates&Rodents	92.3
12	Boreoeutheria	Placental mammals	100.0
13	Eutheria	Placental mammals	104.2
14	Theria	Marsupials&Placentals	162.6
15	Mammalia	Mammals	167.4
16	Amniota	Amniotes	296.0

Table A.8: Sites under episodic positive selection in ARC according to branch-site MEME modelling, number of branches for a site on the path from root to human with support for positive selection (EBF $\geqslant$ 3); All sites with at least one branch indicating positive selection listed, only ones with asterisk were found significantly positively selected by site-by-site likelihood ratio test.

N	Aminoacid	log-LR	LRT p-value	MEME +branches	most recent +branch
2	E	1.60	0.206	1	Mammalia(167.4)
21	Q	0.51	0.477	2	Hominidae(15.7)
23	A	3.21	0.073	1	Mammalia(167.4)
47	R	5.67	0.017*	0	
51	A	0.07	0.785	2	Hominidae(15.7)
82	S	0.03	0.861	1	Catarrhini(29)
84	S	0.17	0.679	1	Mammalia(167.4)
150	V	10.38	0.001*	0	
158	Y	0.16	0.692	2	Theria(162.6)
170	S	0.37	0.545	1	Mammalia(167.4)
179	A	0.05	0.815	1	Theria(162.6)
194	Y	1.10	0.295	1	Mammalia(167.4)
198	V	1.91	0.167	6	Catarrhini(29)
199	P	3.99	0.046*	0	
206	S	1.99	0.159	1	Mammalia(167.4)
208	G	0.28	0.598	1	Mammalia(167.4)
234	S	0.09	0.758	1	Boreoeutheria(100)
256	F	1.67	0.197	1	Boreoeutheria(100)
359	L	1.51	0.22	1	Mammalia(167.4)
367	G	0.11	0.742	1	Mammalia(167.4)
368	P	2.69	0.101	1	Boreoeutheria(100)
370	L	0.16	0.687	3	Hominidae(15.7)
371	P	6.34	0.012*	2	Homo Sapiens(0)
384	N	0.35	0.552	2	Similiformes(42.6)

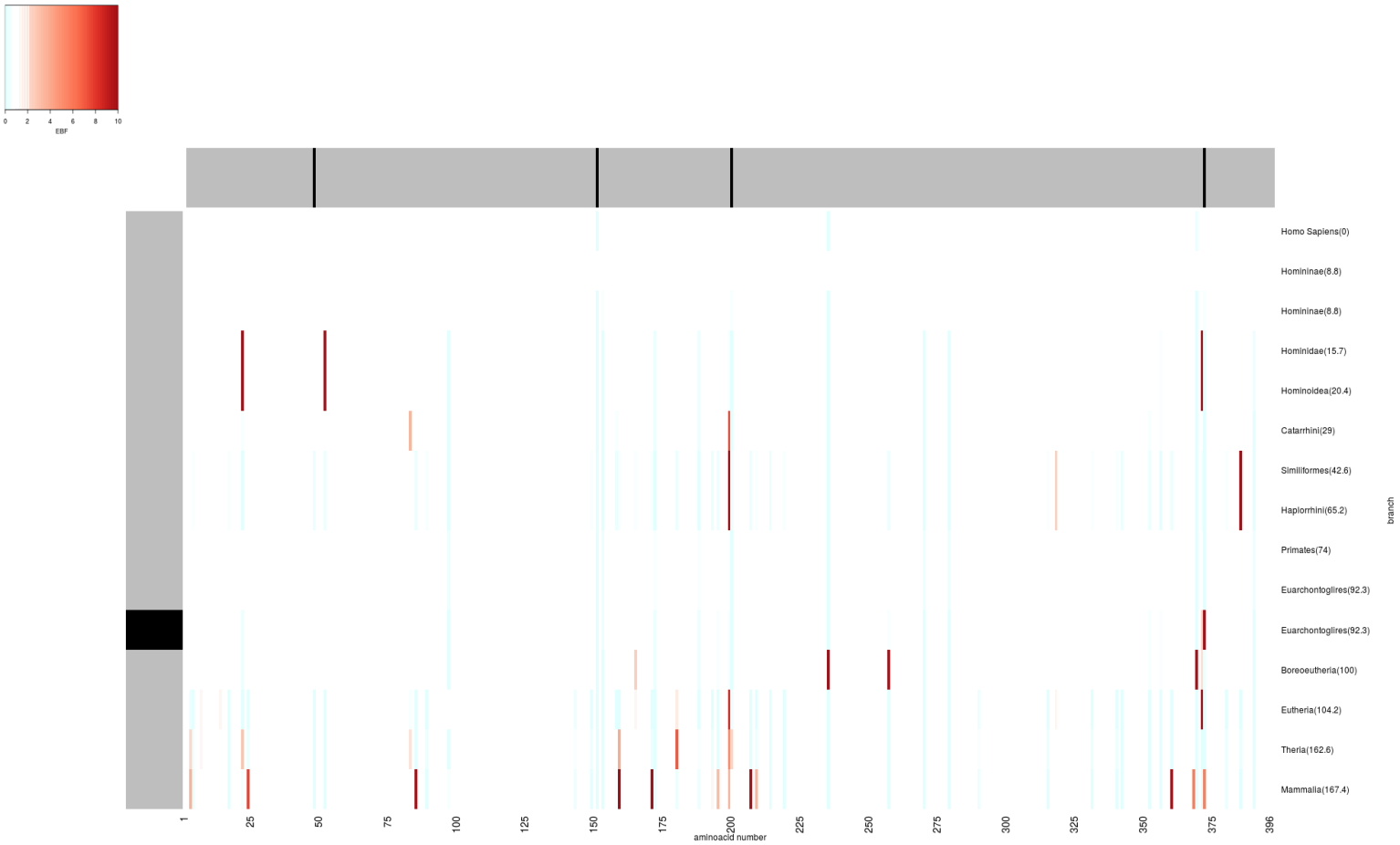


Figure A.5: Spatiotemporal results summary for *Arc*. Vertical grey bar highlights branches from human-root path identified as positive in branch-by-branch analysis (aBSREL),these are annoatated with divergence points and their time estimates (mya) according to Ensembl (See also Table 3.2 and Figure 5.11Horizonal grey bar highlights sites identified as positive in site-by-site analysis (MEME). Cells of the heatmap area represent Bayes Factor in favour of localised episodic positive selection of a (site, branch) tuple (MEME). Red - highly positive, blue - highly negative.

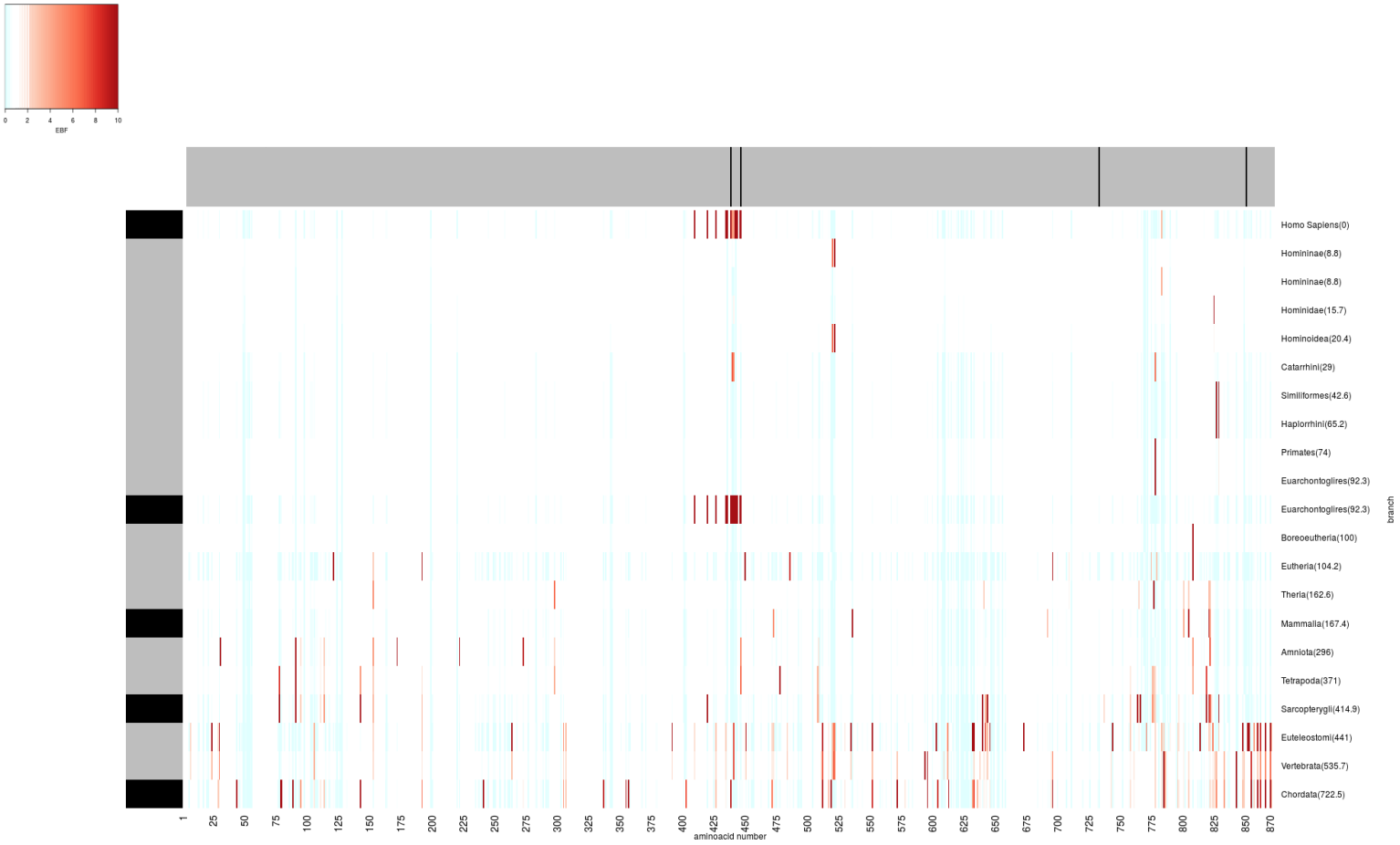


Figure A.6: Spatiotemporal results summary for *Dynamin-2(DNM2)*. Vertical grey bar highlights branches from human-root path identified as positive in branch-by-branch analysis (aBSREL),these are annoatated with divergence points and their time estimates (mya) according to Ensembl (See also Table 3.2 and Figure 5.11Horizonal grey bar highlights sites identified as positive in site-by-site analysis (MEME). Cells of the heatmap area represent Bayes Factor in favour of localised episodic positive selection of a (site, branch) tuple (MEME). Red - highly positive, blue - highly negative.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of March 6, 2018 (`classicthesis` version 4.0).